

# Seminar Large-scale Data Engineering (LDE)

## 02 Scientific Reading and Writing

**Dr.-Ing. Patrick Damme**

Technische Universität Berlin

Berlin Institute for the Foundations of Learning and Data

Big Data Engineering (DAMS Lab)



Last update: Apr 28, 2024

[Credit: Based on “Introduction to Scientific Writing”/  
”02 Scientific Reading and Writing” by Matthias Boehm  
(TU Graz, winter 2021/22)]



# Announcements/Org



- **Hybrid Setting with Optional Attendance**

- In-person in TEL 811 MAR 0.003
- Virtual via zoom

<https://tu-berlin.zoom.us/j/67376691490?pwd=NmlvWTM5VUVWRjU0UGI2bXhBVkxzQT09>



- **Reminder: Selection of Seminar and Project Topics Due Apr 29, 23:59 (TODAY)**

- **Polls in the ISIS course**
- **Seminar:** 5 preferred topics/papers
- **Project:** 5 preferred topics + preference on team/individual work + optionally team members

# Agenda

- **Scientific Reading**
- **Scientific Writing**

Scientific Writing skills can only be learned hands on, and incrementally improved with experience



# Scientific Reading

In Computer Science (Data Management)

# Obtaining the Full Text of a Paper



- If you know the title[, author, venue, year] of a paper

- Use search engines like **DBLP** (<https://dblp.uni-trier.de/>) or **Google Scholar** (<https://scholar.google.com/>)
- Make sure to select the **right version** of the paper

The screenshot shows the DBLP (computer science bibliography) search results for the paper "Compressed Linear Algebra for Large-Scale Machine Learning". The search results are organized by year, with three matches found. The most recent match is from 2019, followed by 2018 and 2016. The 2016 entry is highlighted, showing the authors Ahmed Elgohary, Matthias Boehm, Peter J. Haas, Frederick R. Reiss, and Berthold Reinwald. The paper title is "Compressed Linear Algebra for Large-Scale Machine Learning" and it was published in "Proc. VLDB Endow. 9(12): 960-971 (2016)". The interface also includes a "paper" section with a "view" button and a "BibTeX entry" section with an "export record" button.

The screenshot shows the Google Scholar search results for the same paper. The search results are organized by article type, with "Articles" selected. The most recent article is from 2016, published in "The VLDB Journal" by Springer. The authors are Ahmed Elgohary, Matthias Boehm, Peter J. Haas, and Frederick R. Reiss. The paper title is "Compressed linear algebra for large-scale machine learning". The interface also includes a "view" button and a "BibTeX entry" section with an "export record" button.

- If the paper is **not open-access**, you typically can access the PDF when you

- are in the university VPN (e.g., ACM Digital Library), or
- log in with your university account (redirect to TUB login page) (e.g., IEEE Xplore)



# Finding Related Work



## ■ Motivation

- Some research areas might be very large (e.g., index structures, compression)
- How do you find relevant scientific papers/theses via **multiple channels**

## ■ Prefer Trustworthy Sources

- Archival publications, awareness of peer-review
- From right communities (e.g., ML systems vs ML algorithms)
- Reputation of website, authors, etc.

## ■ Recap: Give Credit

- Cite broadly, **give credit to inspiring ideas**, create connections
- Honestly acknowledge **limitations of your approach**

# Finding Related Work, cont.



## ■ By Venue/Year

- Start off top-tier conferences/journals and find latest work
- E.g., SIGMOD, PVLDB, CIDR, ICDE, EDBT, CIKM, ...
- These papers' related work should provide a good categorization and discussion of related work → [recursive lookup](#)

## ■ By Author

- Sometimes there are well-known experts in a certain sub-area
- Find author publications via DBLP and other libraries



# Finding Related Work, cont.



## ■ By References

- **Backwards** (papers published before)

### 8. REFERENCES

- [1] M. Abadi et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *CoRR*, 2016.
- [2] A. Alexandrov et al. The Stratosphere Platform for Big Data Analytics. *VLDB J.*, 23(6), 2014.
- [3] A. Ashari et al. An Efficient Two-Dimensional Blocking Strategy for Sparse Matrix-Vector Multiplication on GPUs. In *ICS (Intl. Conf. on Supercomputing)*, 2014.
- [4] A. Ashari et al. On Optimizing Machine Learning Workloads via Kernel Fusion. In *PPoPP (Principles and Practice of Parallel Programming)*, 2015.
- [5] M. A. Bassiouni. Data Compression in Scientific and Statistical Databases. *TSE (Trans. SW Eng.)*, 11(10), 1985.
- [6] N. Bell and M. Garland. Implementing Sparse

- & **forwards** (papers published after the given paper)

The screenshot shows a Google Scholar search for "compressed linear algebra". The search bar contains the text "compressed linear algebra" and a search icon. Below the search bar, it says "Scholar About 143.000 results (0,09 sec)". The search results list a paper titled "Compressed linear algebra for large-scale machine learning" by A. Elgohary, M. Boehm, P.J. Haas, and F.R. Reiss, published in the Proceedings of the ... in 2016. The paper is available as a PDF from iisc.ac.in. The search results also show a "Cited by 85" link, which is highlighted with a red box. Other links include "Save", "Cite", "Related articles", and "All 12 versions". The Google Scholar logo is visible in the bottom right corner of the screenshot.

## ■ By Keywords

- Broad survey of other related work, to augment the bias of the year/venue/author approach
- Think of possible synonyms (e.g., “extensible”, “extendable”, “customizable”, ...)



# Types of Reading



## ■ Skimming

- **Goal:** understand what the paper/thesis is about, judge relevance
- Read abstract, and optionally introduction
- Scan paper (sections/subsections, structure, figures)

What?

## ■ Understanding

- **Goal:** understand how the presented approach accomplishes the paper's goals
- #1 Skimming (see above)
- #2 Read the whole paper sequentially, add **notes/annotations**

How?

## ■ Reviewing

- **Goal:** evaluate potential impact, and limitations
- #1 Skimming (see above)
- #2 Understanding (see above) + strengths and weaknesses
- #3 Write summary, strong/weak points, detailed comments, constructive feedback, overall recommendation ([strong/weak] accept/reject)

Good enough?

How to improve?

# Process of Reading – Skimming/Understanding



- **Abstract and Structure**
- **#1 Partial Reading** (mostly skimming)
  - Read into each paragraph until you get what it's about
  - 1<sup>st</sup> sentence/label: **topic sentence**
- **#2 Fast Reading**
  - Normal reading vs **reading w/o vocalization**
  - Avoid need for rereading text
    - Back/forward references,
    - Misplacement after distractions
    - Rereading due to lack of understanding

➔ **Read according to your goals of reading**

# Process of Reading – Understanding/Evaluation



## ■ Skepticism

- Critical reading is important for **understanding** and **evaluation**
- **#1** Start open-minded, listen to arguments and trust provided evidence
- **#2 Don't accept** superficial, contradictory, or unproven claims
- **#3** If there are problems, which **constructive feedback** could you give or how could the problems be addressed?

## ■ Questions to Ask Yourself?

- What is the problem? Is it a real or artificial problem?
- How would you solve the problem yourself?
- How is the paper solving the problem?
- Is this the simplest approach that yields these results (justified complexity)
- Are there limitations that are not covered by the paper?
- Is there existing work that already addresses the same problem?

# Proofreading Your Own Paper



## ■ #1 Read Slowly & Carefully

- **Problem:** Brain interpolates between words
- Awareness of **common syntactic issues** (the the, missing/wrong articles, adapt/adopt)
- Awareness of **common semantic issues** (missing reference, inconsistent / no logical consequence)

→ Read out loud  
→ Use PDF-to-Speech

## ■ #2 Read Fully

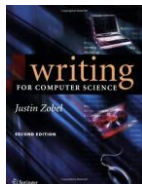
- **Read and annotate issue**, don't fix immediately (destroys the flow)
- Take annotated document and fix issues

## ■ #3 Ask Big Questions

- **Pitfall:** Being **overly focused on syntactic/local issues**
- Is the overall idea clearly communicated and does it make sense?
- Are there missing pieces, missing experiments, missing related work?

# Scientific Writing

## In Computer Science (Data Management)

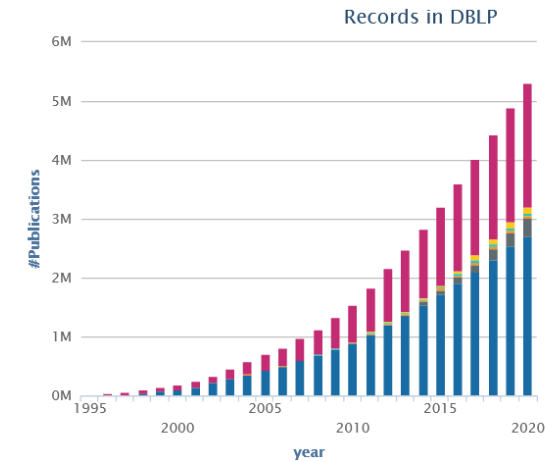


[Justin Zobel: Writing for Computer Science,  
2nd ed. Springer 2004, ISBN 978-1-85233-802-2]

# Recap: Writing the Paper



- **Know Your Audience**
- **Get Your Workflow in Order / Incremental Paper Drafts**
- **Mindset: Quality over Quantity**
  - Aim for top-tier conferences/journals (act as filter)
  - Make the paper useful for others (ideas, evidence, code)



- **Make the Paper Easy to Read**
- **Present Your Work with appropriate Structure, Writing Style, and Formatting**

# Recap: Prototypical Structure of a Scientific Paper



## ■ Sections and Subsections

- Abstract → short overview of problem and solution (part of meta data)
- Introduction → context, problem, contributions
- Background / Preliminaries → necessary background for understanding
- Main Part → your technical core contributions
- Main Part 2
- Experiments → setting, micro benchmarks, end-to-end benchmarks
- Related Work → areas of related work, differences to your own work
- Conclusions → summary, conclusions, and future work
- Acknowledgments → funding agencies, helpful people beyond co-authors
- References → list of other works referenced throughout the paper
- (Appendix) → any additional contents (e.g., proves of theorems, more results)

→ 01 Structure of Scientific Papers

## ■ Recommendations

- Avoid sections with only one subsection (e.g., 2 and 2.1)
- Avoid more than two or at most three nesting levels
- Clearly separate motivation/background from your own work

# Formatting Elements Facilitating Structure

## ■ Goal: Easy Skimming

- Guide the readers' attention

## ■ Paragraph Labels

- `\paragraph{...}` ...

**Data Structure:** The MNC sketch  $\mathbf{h}_A$  of an  $m \times n$  matrix  $A$  comprises the following information, where we use  $\mathbf{h}$  as a shorthand whenever the context is clear.

## ■ Bullet Lists

- `begin{itemize} ... \item ... \end{itemize}`
- `begin{enumerate} ... \item ... \end{enumerate}`

shorthand whenever the context is clear.

- **Row/Column NNZs:** Count vectors  $\mathbf{h}^r = \text{rowSums}(A \neq 0)$  and  $\mathbf{h}^c = \text{colSums}(A \neq 0)$  indicate the NNZs per row and column, where  $h_i^r$  is the count of the  $i$ th row.
- **Extended Row/Column NNZs:** Count vectors  $\mathbf{h}^{er} = \text{rowSums}((A \neq 0) \cdot (\mathbf{h}^c = 1))$  and  $\mathbf{h}^{ec} = \text{colSums}((A \neq 0) \cdot (\mathbf{h}^r = 1))$  indicate the NNZs per row/column that appear in columns/rows with a single non-zero.



Figure 2: Accuracy/Efficiency Goal of the MNC Sketch.

## ■ Figures and Tables

- Should be self-explanatory
- Captions above/below

Table 1: Analysis of Existing Sparsity Estimators.

Estimator	Space	Time	% Bias
MetaAC $E_{ac}$	$O(1)$	$O(1)$	✓
MetaWC $E_{wc}$	$O(1)$	$O(1)$	✓ $\overline{sc}$
Bitset $E_{bitset}$	$O(mn + nl + ml)$	$O(mnl)$	✓
DMap $E_{dm}$	$O(\frac{mn+nl+ml}{p})$	$O(\frac{mnl}{p^2})$	✓
Sample $E_{smp}$	$O( S )$	$O( S (m+l))$	$\overline{sc}$
LGraph $E_{lgraph}$	$O(rd + nnz(A, B))$	$O(r(d + nnz(A, B)))$	✓
MNC $E_{mnc}$	$O(d)$	$O(d + nnz(A, B))$	✓

## ■ Theorem/Definition/Example

- Refine **theorem** environment as needed

**THEOREM 3.1.** Given MNC sketches  $\mathbf{h}_A$  and  $\mathbf{h}_B$  for matrices  $A$  and  $B$ , the output sparsity  $sc$  of the matrix product  $C = AB$  can be exactly computed under the assumptions A1 and A2 via a dot product of  $\mathbf{h}_A^c$  and  $\mathbf{h}_B^r$ :

## ■ Algorithms/Pseudo-Code

- Can be clearer than text, but not always
- Carefully select the right level of abstraction

Algorithm 1 MNC Sparsity Estimation

**Input:** MNC sketches  $\mathbf{h}_A$  and  $\mathbf{h}_B$  for matrices  $A$  and  $B$   
**Output:** Output sparsity  $sc$

```

1: // a) basic and extended sparsity estimation, incl upper bound
2: if  $\max(h_A^c) \leq 1 \vee \max(h_B^r) \leq 1$  then // see Theorem 3.1
3:    $nnz \leftarrow h_A^c \cdot h_B^r$ 
4: else if exists( $h_A^{ec}$ )  $\vee$  exists( $h_B^{er}$ ) then // extended NNZ counts
5:    $nnz \leftarrow h_A^{ec} \cdot h_B^r + (h_A^c - h_A^{ec}) \cdot h_B^{er}$  // exact fraction
6:    $p \leftarrow (nnz(h_A^c) - |h_A^c|) \cdot (nnz(h_B^r) - |h_B^r|) // \#cells$ 
7:    $nnz \leftarrow nnz + E_{dm}(h_A^c - h_A^{ec}, h_B^r - h_B^{er}, p)$  // generic rest
8: else // generic fallback estimate
9:    $p \leftarrow nnz(h_A^c) \cdot nnz(h_B^r) // \#cells$ 
10:   $nnz \leftarrow E_{dm}(h_A^c, h_B^r, p) - p$ 
11: // b) apply lower bound, see Theorem 3.2
12:  $nnz \leftarrow \max(nnz, |h_A^c| > n/2 \cdot |h_B^r| > n/2)$  // lower bound
13: return  $sc \leftarrow nnz/(ml)$ 

```

## ■ Refer to All Figures, Tables, Algorithms in the Text & Place Them Close to the Text



- **Goal: Clear, Easy-to-Read Writing**

- Avoid unnecessary formalism → as simple as possible

- **Formal Language**

- Avoid contractions (“can’t”, “aren’t”, ...)
- No colloquial or slang words

- **Prefer Active Voice**

- Easier to understand, shorter, more interesting
- Use “we” over “I”
- Don’t directly address the reader (no “you”)



In this section, the background and motivation for compressed linear algebra is introduced.



In this section, we provide the background and motivation for compressed linear algebra.

- **Prefer Present Tense**

- Most content of a research paper can be described in present
- Exceptions: user studies, (specific experimental setup), related work

### ■ Variation

- Diversity (structure, length of sentences/paragraphs, choice of words, sentence beginning) helps keeping the reader's attention

✘ The system of rational numbers is incomplete. This was discovered 2000 years ago by the Greeks. The problem arises in squares with sides of unit length. The length of the diagonals of these squares is irrational. This discovery was a serious blow to the Greek mathematicians.

✔ The Greeks discovered 2000 years ago that the system of rational numbers is incomplete. The problem is that some quantities, such as the length of the diagonal of a square with unit sides, are irrational. This discovery was a serious blow to the Greek mathematicians.

### ■ Use of References

- Use `\cite{key1,key2}` for multiple sources
- **Don't use references as nouns**
- **Prefer primary sources**
- Use "et al." for three or more authors

✘ Later, [40] investigated query processing on heavyweight Huffman coding schemes,

✔ Later, Raman and Swart investigated query processing on heavyweight Huffman coding schemes [40],

# Writing Style, cont.



- **Singular/Plural and Articles**

- Plural allows to drop articles

employ general-purpose  
compression techniques

employ **a** general-purpose  
compression technique



- **Guarded Spaces**

- Use guarded spaces for references that should not appear on a new line

Figure~\ref{fig:exp1}

- **Clear References**

- Make sure there are no unclear “it” or “this” references
- Add descriptive nouns

Each entry  $q_i$  can be expressed over columns as  $q_i = v^T X_i$ . We rewrite **this** in [...]



Each entry  $q_i$  can be expressed over columns as  $q_i = v^T X_i$ . We rewrite **this multiplication** in [...]



- **Capitalize Titles and Names**

- Titles: capitalize meaning-carrying words
- Names: capitalize, e.g., Bayesian, Euclidean
- References like Figure 1, Table 2, Section 3, Chapter 4, Equation 5 are names as well

SliceLine: Fast, Linear-Algebra-based Slice Finding  
for ML Model Debugging

Svetlana Sagadeeva\*  
Graz University of Technology

Matthias Boehm  
Graz University of Technology

Figure~\ref{fig:exp1}  
Equation~\eqref{eq:e1}



## ■ Commas

- Whenever a pause is appropriate, or required to avoid ambiguity

When using disk[, ] tree algorithms were found to be particularly poor.

A woman without her man is nothing.  
A woman: without her, man is nothing.

- **Lists:** red, blue, black, and white (Oxford/serial comma)
- **Special sentence start:** However, Hence, Therefore, In this paper,

## ■ Semicolons

- Divide a long sentence into sub-sentences, or separation for emphasis
- Lists with sublists

We use index structures like b-trees, tries, and hash tables; as well as compression techniques like run-length encoding, dictionary encoding, and null suppression.

## ■ Exclamations

- Avoid exclamation marks! Never use more than one!!

# Diversity and Inclusion

[Credit: <https://dbdni.github.io/>]



## ▪ Diversity, “the who”

- Individuals from a wide variety of backgrounds and experience, different viewpoints/reasoning/approaches
- **Different cultures:** e.g., use names from variety of languages, cultures, nationalities (not just Alice and Bob)
- **Differences in figures:** e.g., people-like icons: use variety of gender, skin color, ability status, ...
- **Gender diversity in pronouns:** use variety of he/she/they, use gender-neutral nouns: “chairman” → “chairperson”

## ▪ Increasing Awareness for D&I

- Meanwhile part of the policies of all/most major publication venues (SIGMOD, VLDB, ICDE, EDBT, ADBIS, ...)
- **D&I issues included in the review form**

## ▪ Inclusion, “the how”

- Environment welcoming and embracing diversity; **avoid** language that furthers the marginalization, stereotyping, erasure of any group of people
- **Implicit assumptions:** “Everyone has a mother and a father.”
- **Oppressive terminology:** e.g.,
  - “master-slave” → “coordinator-worker”
  - “orphaned object” → “unreferenced object”
  - “blacklist/whitelist” → “blocklist/allowlist”
- **Marginalization of under-represented groups:** e.g., “The Gender attribute is either Male or Female.”
- **Lack of accessibility:** e.g., color alone to convey info in a plot → use patterns, symbols, textures, etc.
- **Stereotyping:** e.g., feminine names or presentations for personal secretary role

# Formatting



## ■ Goal: Emphasize Quality of Contents with Quality of Visual Presentation

- A carelessly formatted paper (layout, figures, fonts, underlining) creates a bad first impression
- Recap: **skimming** and **anchoring**

“The paper’s approach is probably equally sloppy”

## ■ Figures

- Use same font and font size as the main text / code in main paper
- Avoid text overlap, too aggressive **colors**

## ■ Orphans and Widows

- Imprecise definition
- Avoid few words per line, single line at next page

**Strength Reduction:** Note that `cumsumprod(X)` uses `cumsumprod(B)n1`—i.e., the last block entry—as part of  $f_{agg}$ . Similarly, for `cumsum(X)`, we could use `cumsum(B)n`. However, this simplifies to `colSums(B)`, which avoids materializing the cumsum output block.

Looks ugly and wastes lots of space

## ■ Text Running over Column Margin (rephrase until it fits)

## ■ Highlighting

- `\emph{...}` (emphasize) over underlining or bold
- `\texttt{...}` or `\verb+...+` for inline code

# Page Limits

## ■ Most Conferences/Journals

- Given predefined template, changes not permitted
- SIGMOD/PVDLB: **12 pages + unlimited references**
- ICDE: 12 pages incl. references

## ■ Avoid Cheating

- Don't change the template, fonts, or margins (at least not too excessively)
- Condensing more text into the paper will make it harder to read

## ■ Carefully Trim Down Draft

- Write unlimited paper, then select, and revise
- Write and revise section by section as you write

## ■ Never Excuse Missing Content by “lack of space”

“Due to the lack of space, we omit [essential details] / [essential experiments]”



[Credit: <https://twitter.com/fadeladib/status/1322646406088347649>]



[Eamonn Keogh: How to do good research, get it published in SIGKDD and get it cited!, KDD 2009]





## Self-Plagiarism (Bad Idea)

- Avoid reusing motivation, introduction, figures, and examples
- Start writing every thesis / paper from scratch (unless thesis summaries/extends previous papers)

## Figure Plagiarism (Bad Idea)

- Never copy figures from other papers, web, etc
- Create all figures yourself, even for surveys (can be based on ideas of existing papers)
- **Exceptions** do exist w/ explicit references

### Efficiently Compiling Efficient Query Plans for Modern Hardware

Thomas Neumann  
Technische Universität München  
Munich, Germany  
neumann@in.tum.de

#### ABSTRACT

As main memory grows, query performance is more and more determined by the raw CPU costs of query processing itself. The classical iterator style query processing technique is very simple and flexible, but shows poor performance on modern CPUs due to lack of locality and frequent instruction mis-predictions. Several techniques the hash-orientated processing or vectorized tuple processing have been proposed in the past to improve this situation, but even these techniques are frequently out-performed by hand-written execution plans. In this work we present a novel compilation strategy that translates a query into compact and efficient machine code using the LLVM compiler framework. By aiming at good code and data locality and predictable branch layout the resulting code frequently yields the performance of hand-written C++ code. We integrated these techniques into the HyPer main memory database system and show that this results in excellent query performance while requiring only modest compilation times.

#### 1. INTRODUCTION

Most database systems translate a given query into an expression in a (physical) algebra, and then start evaluating

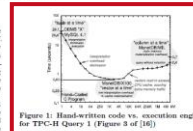


Figure 1: Hand-written code vs. execution engine for TPC-H Query 1. (Figure 3 of [26])

CPUs. Thus, this model often results in poor code locality and complex look-uping. This can be seen by considering a simple table scan over a compressed relation. As the tuples must be probed one at a time, the table scan operator has to remember where in the compressed stream the current tuple is and jump to the corresponding decompression code when asked for the next tuple. These observations have led some modern systems to a

## Plagiarisms (Really Bad Idea)

- Never copy figures or text from other peoples work and claim its yours (slight rewording does not change that)
- For archival scientific publications, there is a high chance it will be detected



## ▪ Example SIGMOD'21

A research paper submitted to SIGMOD 2021 **cannot be under review** for any other publishing forum or presentation venue, including conferences, workshops, and journals, during the time it is being considered for SIGMOD. Furthermore, after you submit a research paper to SIGMOD, you must **await the response** from SIGMOD and only re-submit elsewhere if your paper is rejected - or withdrawn at your request - from SIGMOD. This restriction applies not only to identical papers but also to papers with a substantial overlap in scientific content and results.

Every research paper submitted to SIGMOD 2021 must present substantial novel research not described in any prior publication. In this context, a **prior publication** is (a) **a paper of five pages** or more presented, or accepted for presentation, at a refereed conference or workshop with proceedings; or (b) **an article published**, or accepted for publication, in a refereed journal. If a SIGMOD 2021 submission has overlap with a prior publication, the submission must cite the prior publication, along with all other relevant published work, following the guidelines in the Anonymity Requirements for Double-Blind Reviewing section below.

# Excursus: Automatic CS Paper Generation



## ▪ SCiGen

- Generates random CS research papers, including graphs and figures
- Uses hand-written context-free grammar
- Test for low-submission standards of conferences
- **Meaningless mix of sentences and technical terms**

[Credit: <https://pdos.csail.mit.edu/archive/scigen>]

## ▪ Generative AI (such as ChatGPT)

- **ACM Policy on Authorship (applies to, e.g., SIGMOD)**
  - **Generative AI tools** may not be authors of publications
  - Using generative AI to create content is **permitted**
  - **But: must be fully disclosed in the work**
  - **Basic word processing systems** (e.g., spelling/grammar corrections) generally allowed, no requirement for disclosure
  - Policy updates expected due to blurring boundaries between generative AI and basic word processing systems

[Credit: <https://www.acm.org/publications/policies/new-acm-policy-on-authorship>]

**LDE seminar and project:  
Use of generative AI  
not allowed**

# Summary and Q&A



- Scientific Reading
- Scientific Writing
  
- Remaining **Questions?**
- Seminar/Project Topic Selection by **Apr 29, 23:59 (TODAY)**
  
- Final Introductory Lecture
  - 03 **Experiments, Reproducibility, and Giving Presentations** [May 06, MAR 0.003]  
Also recommendable for participants taking only the project