

CSCI 49395 – Fall 2026

Big Data Analytics and Applied Privacy

Mohammad Hossein Manshaei

www.bdap-hunter.com

Mohammad.Manshaei@hunter.cuny.edu



HUNTER
The City University of New York

Course Projects

How to select research papers and develop a course project?

Contents

- Choosing an Appropriate and Feasible Research Topic
- Finding Relevant Papers and Evaluating Their Quality
- Reading a Scientific Paper and Extracting Its Main Contributions
- Applying the Process to an Example Research Paper
- Planning the Project's Midterm and Final Milestones

Choosing an Appropriate Research Topic

- Address a clear and meaningful **research problem**
- Relate to Big Data Systems, AI and Analytics, or Privacy
- Be supported by relevant scientific papers
- Include an **accessible real-world dataset**
- Require **implementation and experimental** evaluation
- Match the **team's technical skills and interests**
- Be achievable within one semester

Appropriate Research Topics

- Select a topic from one or more course areas:
 - **Big Data Systems**
Spark · Streaming · NoSQL · Cloud · Data Lakes
 - **AI & Analytics**
Machine Learning · LLMs · Distributed Training · Multi-Agent Systems
 - **Privacy & Security**
Privacy Attacks · Anonymization · Differential Privacy · Federated Learning

Best choice: A focused topic connecting multiple course areas.

Contents

- Choosing an Appropriate and Feasible Research Topic
- Finding Relevant Papers and Evaluating Their Quality
- Reading a Scientific Paper and Extracting Its Main Contributions
- Applying the Process to an Example Research Paper
- Planning the Project's Midterm and Final Milestones

How and Where to Search?

- Start by combining keywords from different course areas:
 - **Big Data:** Spark, Streaming, NoSQL, Cloud, Data Lakes, Scalability
 - **AI & Analytics:** Machine learning, LLMs, Distributed Training, Multi-Agent Systems
 - **Privacy:** Differential privacy, Federated Learning, Anonymization, Membership Inference, Data Leakage
- **Where to search:**
 - Google Scholar, Semantic Scholar, ACM Digital Library, IEEE Xplore, DBLP, and Conference websites

Search broadly first, then narrow the topic based on the papers, datasets, and implementations you find.

Search Strategically

- Start Broad, Then Narrow
 - Search for the general research topic
 - Identify common technical terms
 - Add the target method, **dataset**, or **application**
 - Limit results by publication **year**
 - Examine **related** and **cited** papers
- Follow the Research Network
 - **References**: What earlier work does the paper build upon?
 - **Citations**: What later work used, evaluated, or criticized it?
 - **Related papers**: What alternative approaches exist?
 - **Authors**: Did they publish follow-up work?

One **strong paper** can lead you to an entire **research direction**.

Is It a Scientific Research Paper?

- Before evaluating quality, confirm the **publication type**:
 - Full conference paper, journal article, workshop paper, short paper or poster, survey or review paper, and technical report or preprint
- Check:
 - Is it peer-reviewed?
 - Was it accepted by a **recognized venue**?
 - Is a published version available?
 - Is the arXiv version different from the final version?

How to Evaluate the Publication Venue

- Is the **venue recognized** in the paper's research area?
- Is it **organized by a reputable professional society**?
 - Are its proceedings published by ACM, IEEE, USENIX, or another established publisher?
- Are papers **reviewed** by researchers in the field?
- Do **respected researchers** regularly publish there?
- Is it indexed by DBLP, ACM, IEEE, or other scholarly databases?

Examples of Relevant Venues

- **Big Data and Data Systems:** SIGMOD, VLDB, ICDE, KDD, IEEE BigData
- **AI and Machine Learning:** NeurIPS, ICML, ICLR, AAAI, IJCAI
- **Security and Privacy:** IEEE Symposium on Security and Privacy, USENIX Security, ACM CCS, NDSS, PETS

Check Citations and Research Impact

Citations can indicate that a paper has influenced later research.

- **Check:**

- Total citation count, publication year, and citation rate relative to the paper's age
- Whether later papers reproduce, extend, or criticize it
- Whether it introduced a commonly used method or dataset

- **Interpret Carefully**

- Older papers usually have more citations
- Recent high-quality papers may have few citations
- Citation counts differ across research areas
- A highly cited paper is not automatically suitable for the project

Check the Quality of the Publication Venue

- **h5-index** — Google Scholar Metrics
 - Measures the recent citation impact of a conference or publication
 - Compare venues within the **same research field**
- **SJR and Quartile** — SCImago
 - Citation-based measure considering the prestige of citing sources
 - Quartiles: **Q1, Q2, Q3, Q4**
 - Primarily useful for journals and indexed proceedings
- **CORE Conference Ranking**
 - **A*** — Flagship conference
 - **A** — Excellent conference
 - **B** — Good conference
 - **C** — Meets minimum conference standards

Inspect the Dataset and Code

- **Dataset**

- Is it publicly available?
- Is its size manageable?
- Are its format and features documented?
- Does it require special permission?
- Does it contain sensitive information?
- Can it be processed with available resources?

- **Implementation**

- Is an official repository available?
- Are installation instructions provided?
- Are dependencies and software versions specified?
- Are pretrained models or checkpoints available?
- Is the repository complete and maintained?

Final Paper-Selection Checklist

- ☐ What research problem does it address?
- ☐ What are its main contributions?
- ☐ Why is it relevant to BDAP?
- ☐ Where was it published?
- ☐ What dataset does it use?
- ☐ Is the dataset accessible?
- ☐ Is the code available?
- ☐ What experiments can be reproduced?
- ☐ What additional analysis could be performed?
- ☐ What resources and technical skills are required?

Contents

- Choosing an Appropriate and Feasible Research Topic
- Finding Relevant Papers and Evaluating Their Quality
- Reading a Scientific Paper and Extracting Its Main Contributions
- Applying the Process to an Example Research Paper
- Planning the Project's Midterm and Final Milestones

Read a Scientific Paper in Three Passes

- **First Pass: Understand the Story**
 - Read the title, abstract, introduction, main figures, and conclusion
 - Identify the **problem, motivation, proposed idea, and main result**
- **Second Pass: Understand the Method**
 - Study the methodology, key components, assumptions, and implementation
 - Explain: **Input → Processing → Proposed Method → Output**
- **Third Pass: Evaluate the Evidence**
 - Examine datasets, baselines, metrics, experiments, and limitations
 - Determine whether the results support the authors' claims

Extract the Main Contributions

- Identify what the paper adds beyond previous work:
 - **Problem** — Introduces or reformulates an important research problem
 - **Method** — Proposes a new algorithm, model, or system
 - **Data** — Introduces a dataset, benchmark, or processing pipeline
 - **Theory** — Provides a formal analysis, proof, or guarantee
 - **Experiments** — Produces new findings through evaluation

Read the Paper Critically

- **Evaluate the Paper**

- Are the datasets, baselines, and metrics appropriate?
- Are the comparisons fair and convincing?
- Are improvements consistent across different settings?
- What are the accuracy, privacy, scalability, and cost tradeoffs?
- What assumptions and limitations remain?

Find Project Opportunities

- **Turn Limitations into Project Ideas**
 - Test another dataset or setting
 - Add a stronger baseline or evaluation metric
 - Study robustness, scalability, or computational cost
 - Investigate privacy leakage or privacy–utility tradeoffs
 - Reproduce selected results and explain any differences

Contents

- Choosing an Appropriate and Feasible Research Topic
- Finding Relevant Papers and Evaluating Their Quality
- Reading a Scientific Paper and Extracting Its Main Contributions
- Applying the Process to an Example Research Paper
- Planning the Project's Midterm and Final Milestones

Let's Work Through an Example!

CAN LLMs KEEP A SECRET? TESTING PRIVACY IMPLICATIONS OF LANGUAGE MODELS VIA CONTEXTUAL INTEGRITY THEORY

Why This Paper Fits the BDAP Project?

Niloofar Mireshghallah^{1*} Hyunwoo Kim^{2*}
Xuhui Zhou³ Yulia Tsvetkov¹ Maarten Sap^{2,3} Reza Shokri⁴ Yejin Choi^{1,2}
¹University of Washington ²Allen Institute for Artificial Intelligence
³Carnegie Mellon University ⁴National University of Singapore
niloofar@cs.washington.edu hyunwook@allenai.org
<https://confaide.github.io>

ABSTRACT

Existing efforts on quantifying privacy implications for large language models (LLMs) solely focus on measuring leakage of training data. In this work, we shed light on the often-overlooked interactive settings where an LLM receives information from multiple sources at inference time and generates an output to be shared with other entities, creating the potential of exposing sensitive input data in inappropriate contexts. In these scenarios, humans naturally uphold privacy by choosing whether or not to disclose information depending on the context. We ask the question “*Can LLMs demonstrate an equivalent discernment and reasoning capability when considering privacy in context?*” We propose CONFAIDE, a benchmark grounded in the theory of contextual integrity and designed to identify critical weaknesses in the privacy reasoning capabilities of instruction-tuned LLMs. CONFAIDE consists of four tiers, gradually increasing in complexity, with the final tier evaluating contextual privacy reasoning and theory of mind capabilities. Our experiments show that even commercial models such as GPT-4 and ChatGPT reveal private information in contexts that humans would not, 39% and 57% of the time, respectively, highlighting the urgent need for a new direction of privacy-preserving approaches as we demonstrate a larger underlying problem stemmed in the models’ lack of reasoning capabilities.

LLMs + Privacy +
Benchmark Data + Experimental Evaluation

Contents

- Choosing an Appropriate and Feasible Research Topic
- Finding Relevant Papers and Evaluating Their Quality
- Reading a Scientific Paper and Extracting Its Main Contributions
- Applying the Process to an Example Research Paper
- Planning the Project's Midterm and Final Milestones

Important Dates

- **Starting Today**
 - Begin discussing project ideas in the **Brightspace Forum**
 - Form a team of **three students**
 - Share potential topics and candidate research papers
 - TAs and I will participate in the discussions and provide feedback and suggestions
 - Identify at least **three closely related research papers**
- **Midterm Project Presentation**
 - Scheduled for **early October** (exact date TBA)
 - Present your research problem, selected papers, dataset, and initial progress
- **Final Project Presentation**
 - Scheduled during the **last week of classes**
 - Present your implementation, experimental results, and analysis