# Besat Kassaie

## Postdoctoral Researcher

Data Systems Group, The Cheriton School of Computer Science, University of Waterloo
Waterloo, ON N2L 3G1

Phone: +1(519)722-5963 | Email: bkassaie@uwaterloo.ca | Nationality: Canadian

Personal Webpage: https://besatkassaie.github.io/

## Research Interests

**Data Science:** Natural Language Processing, Information Extraction, Unstructured Data Quality, Data Discovery
**Information Retrieval:** Mathematical Information Retrieval
**Applied Machine Learning:** Health Informatics, Large Language Models for Data Management

## Education

The Cheriton School of Computer Science. University of Waterloo. Canada

PhD in Computer Science, 2023.

PhD Thesis: Update-Aware Information Extraction.

Thesis Advisor: Prof. Frank William Tompa.

Committee Members: Prof. Doan from University of Wisconsin; Prof. Ilyas, Prof. Kerschbaum, and Prof. Gurfinkel from University of Waterloo.

Science and Research Branch. Azad University. Iran

MENG in Software Engineering, 2009.

Science and Research Branch of Azad University is among the top universities for graduate studies in Iran.

Ranked in the top 1% of nationwide applicants (23/50000)

Masters Thesis: Application of Textual Corpus in Ontology Matching.

Thesis Advisor: Prof. Amir Masoud Rahmani.

Iran University of Science and Technology. Iran

BENG in Software Engineering, 2005.

Iran University of Science and Technology is among the top five universities in Iran.

Ranked in the top 1% of nationwide applicants (243/500000).

Undergrad Research Advisor: Prof. Mohsen Sharifi.

## Professional  Memberships

Association for Computing Machinery (ACM).

## Languages

| | |
|---|---|
| English: Fluent | Turkish: Native |
| Farsi: Native | Arabic: Basic |

## Research and Industrial Experience

Current     Postdoctoral Researcher. The Cheriton School of Computer Science, University of Waterloo, Canada.

I am involved in research focused on data intelligence and data lakes. I assist with the drafting of grant proposals and academic papers.

2024-01     Visiting Researcher. The Cheriton School of Computer Science, University of Waterloo, Canada.
2024-09

I served as a co-investigator on the math retrieval project. My responsibility involved devising an effective query from a text that contains both mathematical formulas and textual content.

2019-2023   Graduate Research Assistant. Innovation Lab, University of Waterloo, Canada.

I was a member of the research team developing a platform that searches documents with text and mathematical content using a pen-based interface.

2015-2019   Graduate Teaching Assistant & Graduate Research Assistant. The Cheriton School of Computer Science, University of Waterloo, Canada.

During my PhD studies, I was a teaching assistant for multiple undergraduate courses in Data Structures & Algorithms, Database Systems, and Introduction to Computer Science. My responsibilities included assisting in recitations, performance assessments, and design of course assignments and projects. For each course, multiple sessions are offered with hundreds of students in each session.

2014        Senior Research Engineer. AGENCY FOR SCIENCE, TECHNOLOGY AND RESEARCH (A*STAR), Singapore.

I was a senior research engineer in a research team working on a project for the Mclaren company. The team was working on developing a machine-learning-based approach to design a sleep quality assessment tool. My responsibilities included: model deployment and evaluation as well as participating in conducting and analyzing sleep quality experiments. A*STAR is a catalyst and enabler of significant research initiatives among the research community in Singapore and beyond.

2013-2014   Information Technology Consultant. ITFORCE, Singapore.

I was a consultant on software engineering projects.

2009-2013   Senior Data Engineer. MCLS, Iran.

I was a Senior Data Engineer in a large-scale data integration project.

## Teaching and Mentorship Experience

Instructor. CS338. Computer Applications in Business: Databases. University of Waterloo. Winter 2025.
One session 100 students and several TAs.

Mentor. The School of Computer Science, University of Waterloo. Fall 2024.
Undergraduate Students:
- Alice Zhao

   Software Engineering @ University of Waterloo
- Jeri Fan

   Computer Science @ University of Waterloo

   Business Administration @ Wilfrid Laurier University
- Fiona Chen

   Computer Science @ University of Waterloo

   Business Administration @ Wilfrid Laurier University

Teaching Assistant. CS338: Computer Applications In Business: Databases. University Of Waterloo. Winter 2017, Winter 2019. Multiple sessions with ~200 students and several TAs each.
Main responsibilities: Design of course assignments and projects, recitation, and performance assessment.

Teaching Assistant. CS348: Introduction to Database Management. University of Waterloo.
Spring 2017, Fall 2017. Multiple sessions with ~250 students and several TAs each.
Main responsibilities: Recitation and performance assessment.

Teaching Assistant. CS240: Data Structures and Data Management. University of Waterloo.
Winter 2016. Multiple sessions with ~300 students and several TAs each.
Main responsibilities: Recitation and performance assessment.

Teaching Assistant. CS115: Introduction to Computer Science. University of Waterloo.
Spring 2015. Multiple sessions with ~300 students and several TAs each.
Main responsibilities: Recitation and performance assessment.

Teaching Assistant. CS 116x: Introduction to Computer Science. University of Waterloo.
Winter 2015. Multiple sessions with ~250 students and several TAs each.
Main responsibilities: Recitation and performance assessment.

## Publications

### JOURNAL PAPERS

P1- Besat Kassaie, Frank Wm. Tompa. 2024. Updatable Extracted Views for Improving Unstructured Data Quality. Submitted to The ACM Journal of Data and Information Quality (JDIQ) .
JDIQ's mission is to publish high quality articles that make a significant and novel contribution to the field of data and information quality.

P2- Besat Kassaie, Frank Wm. Tompa. 2023. Autonomously Computable Information Extraction. In Proceedings of the VLDB Endowment, 16, 10, pp. 2431-2443. 13 pages.
PVLDB is a major journal for high-impact research on data management with an acceptance rate of 18.6%.
🔴REC https://www.youtube.com/watch?v=uLpGBwbsTMU

P3- Besat Kassaie, Elizabeth L. Irving, and Frank Wm. Tompa. 2021. Computer-Assisted Cohort Identification in Practice. In ACM Transactions on Computing for Healthcare (HEALTH) 3, 2, Article 17. 21 pages.
HEALTH is a multi-disciplinary journal for the publication of high-quality original research papers that have scientific and technological results pertaining to how computing is improving healthcare.

### PEER-REVIEWED CONFERENCE

P4- Besat Kassaie, Andrew Kane, Frank Wm. Tompa. 2024. Exploiting Query Reformulation and Reciprocal Rank Fusion in Natural Language Mathematical Search. Under Preparation. 10 pages.

P5- Yin Ki Ng, Dallas J Fraser, Besat Kassaie, Frank Wm. Tompa. 2021. Dowsing for Math Answers. In Proceedings of CLEF 2021. 12 pages.
CLEF promotes research, innovation, and development of information access systems with an emphasis on multilingual and multimodal information with various levels of structure.
🏆 Best of Labs Paper.

P6- Besat Kassaie, and Frank Wm. Tompa. 2020. A Framework for Extracted View Maintenance. In Proceedings of the ACM Symposium on Document Engineering 2020. Association for Computing Machinery. New York, USA, Article 14. 4 pages.
The ACM Symposium on Document Engineering publishes original research papers that focus on the design, implementation, development, management, use and evaluation of advanced systems where documents and document collections play a key role with an acceptance rate of 35%.

P7- Besat Kassaie and Frank Wm. Tompa. 2019. Predictable and Consistent Information Extraction. In Proceedings of the ACM Symposium on Document Engineering 2019. Association for Computing Machinery, New York, USA, Article 1414, 1–10. 10 pages.

### THESES

P8- Besat Kassaie. 2023. Update-Aware Information Extraction. Doctoral Thesis. The Cheriton School of Computer Science, University of Waterloo. Waterloo, Ontario, Canada. 134 pages.

P9- Besat Kassaie. 2009. Application of Textual Corpus in Ontology Matching. Masters Thesis. Science and Research Branch, Azad University. Tehran, Iran. 90 pages.

### WORKSHOP PAPERS

P10- Yin Ki Ng, Dallas J Fraser, Besat Kassaie, Frank Wm. Tompa. 2021. Dowsing for Answers to Math Questions: Ongoing Viability of Traditional MathIR. In CLEF 2021 Working Notes. http://ceur-ws.org/Vol-2936/. 19 pages.

P11- Yin Ki NG, Dallas J. Fraser, Besat Kassaie, George Labahn, Mirette S.Marzouk, Frank Wm. Tompa, and Kevin Wang. 2020. Dowsing for Answers with Tangent-L. In CLEF 2020 Working Notes. http://ceur-ws.org/Vol-2696/. 39 pages.

### TECHNICAL REPORTS

P12- Besat Kassaie and Frank Wm. Tompa. 2020. Detecting Opportunities for Differential Maintenance of Extracted Views. arXiv:2007.01973. 19 pages.

P13- Besat Kassaie. 2017. SPARQL over GraphX. arXiv:1701.03091. 11 pages.

## Selected Research Projects

### Novelty-Aware Unionable Table Search (ongoing)

Novelty-Aware Unionable Table Search is an innovative approach designed to identify unionable tables within data lakes while preserving the novelty of the results. This research focuses on curating high-quality, diverse datasets that avoid redundancy and introduce unique insights. By ensuring both unionability and result novelty, the methodology enhances data-driven decision-making processes.

#### OBJECTIVES
Under development to demonstrate the practicality of Novelty-Aware Unionable Table Search.

#### RESPONSIBILITIES
I am the designer and developer of the system.

#### UTILIZED TECHNOLOGIES
Python.

#### FUNDING
NSERC, the Natural Sciences and Engineering Research Council of Canada.

### Differential Maintenance Engine: DME

DME identifies documents' updates that can be autonomously applied on extracted relations. DME statically analyzes an update expression and an extraction program, expressed as a core AQL query, to test sufficient conditions for being a pseudo-irrelevant update. If the input program passes the test, the extracted view content is updated by running a shift algorithm. If it does not pass the test the extractor needs to be executed from scratch.

**OBJECTIVES**

Developed to demonstrate the practicality of the differential maintenance approach proposed in my PhD thesis.

**RESPONSIBILITIES**

I was the sole designer and developer of the engine. Also, I have designed realistic extractors for the DBLP dataset to show the applicability of the proposed optimization strategy in practice.

**UTILIZED TECHNOLOGIES**

Java/Scala Programming Languages. AQL/JAPE Information Extraction Languages.

**ACHIEVEMENTS**

The experiments conducted using DME have been published in the VLDB journal.

**CODE REPOSITORY**

github.com/Besatkassaie/Differential-Maintenance-Engine

**FUNDING**

NSERC, the Natural Sciences and Engineering Research Council of Canada.

## Updatable Views Verifier: U2V

U2V, short for **U**pdate-**to**-**V**iew, serves as a tool for assessing the updatability of an extracted view presented as a JAPE program. U2V determines whether the view is updatable, i.e., whether it qualifies as a stable extractor. This process unfolds in two key steps: a formal representation of the extractor is created; U2V verifies four sufficient conditions for stability on the formal representation. If a program is determined to be stable, any domain preserving update to the view can be safely propagated to the source documents.

**OBJECTIVES**

Developed to demonstrate the practicality of the update translation mechanism proposed in the PhD thesis.

**RESPONSIBILITIES**

I was the sole designer and developer of the engine.

**UTILIZED TECHNOLOGIES**

Java/Scala Programming Languages. AQL/JAPE Information Extraction Languages.

**CODE REPOSITORY**

git.uwaterloo.ca/bkassaie/updatableviews

**FUNDING**

NSERC, the Natural Sciences and Engineering Research Council of Canada.

## Iterative Query Refinement System: IQREF

IQRef is an alternative approach to active learning for expert-in-the-loop cohort identification. IQRef uses a hold-out sample for evaluating cohort selection criteria and deploys an adaptive data analysis technique to prevent overfitting to the hold-out data. IQRef operates in two modes: Exploration in which the learner builds queries with high precision; Integration in which the learner combines multiple queries to achieve high recall.

### OBJECTIVES

Developed to assist the University of Waterloo's optometry researchers to conduct medical studies using EHR.

### RESPONSIBILITIES

I was the sole designer and developer of the system.

### UTILIZED TECHNOLOGIES

Java/Python Programming Languages. Apache Lucene Search Engine.

### ACHIEVEMENTS

IQRef is published in the ACM HEALTH journal.

### FUNDING

NSERC, the Natural Sciences and Engineering Research Council of Canada.

## Math-Aware Search Engine

A Math Retrieval system that is developed by the University of Waterloo's researchers. The engine is part of BrushSearch, a platform for searching documents with text and mathematical content using a pen-based interface: https://www.scg.uwaterloo.ca/brushsearch/.

### OBJECTIVES

Developed to search and retrieve documents that include mathematical content for given math formulas and terms.

### RESPONSIBILITIES

I implemented many features of the search engine, including SLT tree conversion to math tokens, extraction of repetition tokens, and formula normalizations. I conducted the study of proximity signals and created the proximity re-ranking run for the 2021 ARQMath lab.

### UTILIZED TECHNOLOGIES

Java/Python Programming Languages. Apache Lucene Search Engine.

## Selected Research Projects (Cont.)

**ACHIEVEMENTS**

The best participant run of the Answer Retrieval task in the ARQMath Lab in years 2020 and 2021; also, the best automatic run of the Formula Retrieval task in the ARQMath Lab in year 2021.

**CODE REPOSITORY**

github.com/kikingo501/MathDowsers-ARQMath

**FUNDING**

NSERC, the Natural Sciences and Engineering Research Council of Canada. Huawei, WATERLOO-HUAWEI Joint Innovation Lab.

## Talks and Presentations

T1- Besat Kassaie. December 2024. Updatable Extracted Views. Symposium oral presentation, Ontario DataBase Day Symposium, University of Waterloo, Waterloo, ON, Canada.

T2- Besat Kassaie. December 2023. Update-Aware Information Extraction. Symposium oral presentation, Ontario DataBase Day Symposium, McMaster University, Hamilton, ON, Canada.

T3- Besat Kassaie. August 2023. Autonomously Computable Information Extraction. Conference oral presentation, VLDB2023 Vancouver, BC, Canada.

T4- Besat Kassaie. August 2023. Autonomously Computable Information Extraction. Poster, VLDB2023 Vancouver, BC, Canada.

T5- Besat Kassaie. March 2023. Autonomously Computable Information Extraction. Departmental seminar, The Cheriton School of Computer Science, University of Waterloo.

T6- Besat Kassaie. October 2020. Introduction to MATH Information Retrieval. Oral presentation, The Cheriton School of Computer Science, University of Waterloo.
I presented our research on mathematical information retrieval to members of WATERLOO-HUAWEI Joint Lab.

T7- Besat Kassaie. Frank Wm Tompa. September 2020. A Framework for Extracted View Maintenance. Conference oral presentation, DocEng2020 San Jose, CA, USA.

T8- Besat Kassaie. May 2019. Predictable and Consistent Information Extraction. Departmental seminar, The Cheriton School of Computer Science, University of Waterloo.

T9- Besat Kassaie. Frank Wm Tompa. September 2019. Predictable and Consistent Information Extraction. Conference oral presentation, DocEng19 Berlin, Germany.

## Talks and Presentations (Cont.)

T10- Besat Kassaie. June 2019. Predictable and Consistent Information Extraction. Poster, The 1st Annual CS-Can Student Symposium, McGill University, Montreal.

I was invited to present my research in a poster session showcasing emerging researchers from every field in computer science and engineering.

T11- Besat Kassaie. September 2018. Applying Differential Privacy to Text. Poster, CPI Event. Waterloo CyberSecurity and Privacy Institute.

I was invited to present my research in a poster session as part of the University of Waterloo's Cybersecurity and Privacy Institute (CPI)'s event.

T12- Besat Kassaie. December 2017. Applying Local Differential Privacy to Text. Departmental seminar, The Cheriton School of Computer Science, University of Waterloo.

## Professional Development and Services

### Professional Services

S1- PC Member. The 18th ACM International Conference on Web Search and Data Mining **WSDM.** 2025.

S2- Journal Reviewer. Springer WWW Journal. 2024.
"Springer WWW is an international, archival, peer-reviewed journal that covers all aspects of the Web."

S3- Conference Reviewer. The 33rd ACM International Conference on Information and Knowledge Management **CIKM.** 2024.

S4- PC Member. The ACM Symposium on Document Engineering **DocEng.** 2024.
"DocEng is the leading international ACM Symposium for researchers, practitioners to explore cutting-edge ideas in the domain of document engineering."

S5- Conference Session Chair. The 49th International Conference on Very Large Databases **VLDB.** 2023.
"VLDB is a premier annual international forum for data management and database researchers, vendors, practitioners, application developers, and users."

S6- Conference Session Chair. The ACM Symposium on Document Engineering **DocEng.** 2021.

S7- Conference Reviewer. The ACM Symposium on Document Engineering **DocEng**. 2017, 2018, 2019, 2020, 2021, and 2024.

S8- Big Data Challenge Reviewer. STEM Fellowship Charity. 2021, and 2023.
"STEM Fellowship is a Canadian registered charity that uses mentorship and experiential learning to equip the next generation of change-makers with indispensable skills in data science."

S9- Journal Reviewer. Springer Applied Sciences Journal. 2020.
"SN Applied Sciences is a open access journal and indexed in Web of Science's Emerging Sources Citation Index (ESCI), SCOPUS, Ei Compendex and DOAJ."

S10- Journal Reviewer. The International Journal on Semantic Web and Information Systems **IJSWIS.** 2017.
"IJSWIS is a journal that publishes high quality original manuscripts in all aspects of Semantic Web."

S11- Conference Reviewer. The International Conference on Data Management Technologies and Applications **DATA.** 2017.
"The purpose of DATA is to bring together researchers interested on databases."

**Developments**

D1- Canadian Celebration of Women in Computing Conference. Participant. York University. Toronto. 2024.
"CAN-CWiC is the premiere Canadian computing conference for young women and non-binary people in technology."

D2- Graduate Student Mentor Training. Computing Research Association (CRA). Participant. 2024.
"This mentor training course helps graduate students to support undergraduate researchers in a culturally responsive manner."

D3- Machine Learning Theory Workshop. University of Waterloo. Participant. 2023.
"A day-long seminar, offered by Vector Institute,  including: research talks by senior members of the community, group discussions and a poster session"

D4- Waterloo Woman in Computer Science Initiative. The Cheriton School of Computer Science. University of Waterloo. Member. 2015-2023.
"Women in Computer Science (WiCS) is dedicated to promoting gender equity in computing by advocating for and supporting women, trans, gender-fluid, gender-queer, and non-binary students enrolled in computer science and related computing programs at the University of Waterloo."

D5- Grad Cohort for Women. Online Event. Participant. 2022.
"At the Grad Cohort for Women, we spent two days interacting with about 20 senior female computing-related researchers and professionals, who  shared pertinent information on graduate school survival skills, as well as more personal information and insights about their experiences."

D6- Deep Learning and Reinforcement Learning Summer School.Toronto. Participant. 2018.
"DLRL is a keystone next-generation offering of the CIFAR Learning in Machines & Brains program and the CIFAR Pan-Canadian AI Strategy's AI4Good Training Program, hosted each year in partnership with Canada's three national AI institutes: Amii in Edmonton, Mila in Montreal and the Vector Institute in Toronto."

D7- A/B Testing at Scale: Accelerating Software Innovation. Tokyo. Japan.  Tutorial Participant. 2017.
"The goal of the tutorial is to teach attendees how to scale experimentation for their teams, products, and companies, leading to better data-driven decisions. We also want to inspire more academic research in the relatively new and rapidly evolving field of online controlled experimentation."

## Awards and Grants

A1- Graduate Student Research Dissemination Award. University of Waterloo. Graduate Studies and Postdoctoral Affairs. 2023. Competitive.

The intention is to encourage graduate students to present their own research at an academic conference and engage in academic dialogue within their field of study and research.

A2- Graduate Studies Conference Funding. University of Waterloo. The Faculty of Mathematics 2023. Competitive.

The intention is to support graduate students while participating in academic travel and international experiences.

A3- The Doctoral Thesis Completion Award. University of Waterloo. The Faculty of Mathematics. 2023. Competitive.

The intention is to assist highly qualified, full-time doctoral students to complete their thesis writing and defense.

A4- Student Travel Grants. The ACM Special Interest Group on Hypertext and the Web. 2019. Competitive.

The grant was for students traveling to attend any SIGWEB Sponsored Conference.

A5- Math Domestic Graduate Student Award. University of Waterloo. 2015, 2016, 2017, 2018. Competitive.

The goal of this award is to support domestic graduate students in the Faculty of Mathematics, who are engaged in research-based (thesis) programs.

A6- Provost Doctoral Entrance Award for Women. University of Waterloo. 2015. Competitive.

The main purpose of this award is to provide outstanding full-time female doctoral students with an entrance scholarship.

A7- Mathematics Graduate Experience Award. University of Waterloo. 2015. Competitive.

This award is intended to provide financial support for full-time graduate students who acquire experience as a Teaching Assistant during the course of their graduate degree program in the University of Waterloo.

## References

**Frank William Tompa (PhD Supervisor)**

Distinguished Professor Emeritus and Adjunct Professor. Data Systems Research Group. The David R. Cheriton School of Computer Science. University of Waterloo. Canada. fwtompa@uwaterloo.ca. (519) 885 2776.

- Awarded the degree **Doctor of Laws**, *honoris causa*, by **Dalhousie University** for "outstanding personal achievements".
- Founder of **OpenText Corporation**, Canada's fourth-largest software company as of 2022.
- Recipient of **Queen Elizabeth II Diamond Jubilee Medal** for his significant contributions to text data and design systems for maintaining large reference texts.
- **ACM Fellow** for his contributions to text-dominated and semi-structured data management.
- Recognized by the University of Waterloo and the City of Waterloo for epitomizing the energy and enterprise that characterize the University of Waterloo through the naming of the road **Frank Tompa Drive**.

**Renée Miller (Postdoctoral Supervisor)**

Professor. Data Systems Research Group. The David R. Cheriton School of Computer Science. University of Waterloo. Canada. rjmiller@uwaterloo.ca.

- **Canada Excellence Research Chair In Data Intelligence.**
- Fellow of the Association for Computing Machinery
- Fellow of the Royal Society of Canada.
- Recipient of **US Presidential Early Career Award**.
- Recipient of **National Science Foundation CAREER Award**.
- Recipient of **Ontario Premier's Research Excellence Award**.
- Recipient of **IBM Faculty Award**.

**AnHai Doan (Academia)**

Professor. UW-MADISON Database Group. Department of Computer Science. USA. University Of Wisconsin. anhaid@gmail.com. (608) 262 9759.

- **Vilas Distinguished Achievement Professor**.
- Recipient of the **ACM Doctoral Dissertation Award.**
- Recipient of the **CAREER Award**.
- Co-authore of a Morgan-Kaufmann textbook **"Principles of Data Integration"**.
- **Gurindar S. Sohi Professor**.

**Elizabeth L Irving (Academia)**

Professor. School of Optometry And Vision Science. University of Waterloo. Canada. elirving@uwaterloo.ca. (519) 888 4567 x45529.

- **Canada Research Chair** in Vision Science.
- Affiliated with the Eye Research Institute of Canada in Toronto.
- Affiliated with the Department of Ophthalmology at the University of Toronto.
- Recipient of **Royal Society of Canada's Alice Wilson Award**.