Akhil Arora

EPFL DLAB

+41 779876871
akhil.arora@epfl.ch
dlab.epfl.ch/people/aarora/
akhil-arora

Summary

Computer scientist with **5 years** of **industry** and **6 years** of **academic research experience** in data and network science, natural language processing, machine learning, information retrieval, and causal inference. DBLP — Google Scholar

Academic Experience

Sep 2018 – Present **Doctoral Researcher**, EPFL | **Formal Collaborator**, Wikimedia Foundation

Developing ML methods to model and improve human knowledge seeking on the Web.

- **LLMNav**, a framework for simulating human website browsing behavior.
- First **privacy-preserving generative model** of human Web browsing behavior.
- **Natural experiment** to measure the **causal effect** of de-orphanization on article visibility in more than **300 language versions** of Wikipedia.
- **Eigenthemes**, the state-of-the-art **unsupervised entity linker** with **700x faster inference** and comparable efficacy to transformer-based alternatives.
- **PARIS+**, a probabilistic model with superior efficacy, **1000x faster inference**, and **10x smaller memory footprint** than neural methods for entity alignment.
- Publications in **EMNLP**, **NAACL**, **SIGIR**, **WSDM**, **VLDB**, and **ICWSM**.

Industry Experience

Jul 2017 – 2018 **Research Scientist**, American Express AI Labs, Bangalore, India

Devised scalable **deep learning algorithms** for the credit card fraud and risk assessment business of American Express leading to **multi-million-dollar impact**.

- BAE, an ensemble of **data bagged Autoencoders**, which **improved** credit card **fraud detection** by **2**% processing $\sim$ **10M** transactions every day.
- NL2SQL, **sequence-to-sequence model** for translating natural language instructions to SQL queries for descriptive analytics.
- TextRisk, an **LSTM-based model** trained on transcribed customer care conversations, which **improved delinquency prediction** performance by **5**%.

Jul 2014 – 2017 **Researcher**, Xerox Research Centre India (XRCI)

Led a team of 10 research scientists and engineers for devising scalable data management algorithms and machine learning models to solve a gamut of complex real-world problems for the Customer-care and Health-care business of Xerox. Highlights:

- GaBiD, a 360° **customer journey** analytics framework with novel models for churn prediction, root-cause identification, and prevention.
- KEO, information extraction framework to construct enterprise knowledge graphs.
- Multiple **patent** applications and publications in **SIGMOD**, **VLDB**, and **WWW**.

Jul 2013 – 2014 **Software Engineer**, Intel Corporation, India

Worked on research problems in security while performing white hat hacking on internal Intel products, security code reviews, assessments, and code assisted penetration. Developed a framework called IronCrow, which was published in **Black Hat** 2014.

Education

2018 – Present **PhD in Computer Science**, *EPFL, Switzerland*
Advisor: *Prof. Robert West*

2011 – 2013 **Masters in Computer Science**, *IIT Kanpur, India*
Advisor: *Prof. Arnab Bhattacharya*

2006 – 2010 **BE in Computer Science**, *The NorthCap University, Gurugram, India*
Advisor: *Prof. Charu Rana*

Technical Skills

Programming	Python, C++, Java, SQL, R, Bash and Shell scripting
Frameworks	PyTorch, Spark, Keras, Boost, Pandas, TensorFlow
ML/NLP skills	Substantial experience training Graph Neural Networks **(GNNs)** and Language Models **(LMs)** on **multilingual Web-scale data**. Experience with Large Language Models **(LLMs)**: **prompting, RAG, and constrained decoding**. Recently, I have also started exploring LoRA fine-tuning of LLMs
Data skills	Experience working with **petabyte-scale** credit-card transaction data at American Express, several hundred **terabytes** of Wikipedia **server logs** with digital traces of human browsing behavior, terabyte-scale **online news** data, text and graph data from **Wikipedia in 300+ languages**, **Wikidata** and NCBI **Knowledge Graphs**, and social networks such as **Twitter** with **millions of nodes** and **billions of edges**
Misc.	MySQL, Elasticsearch, GraphX, LaTeX, OpenCV, Matlab, GNU Octave, NodeJs, MongoDB, Dockers, Heroku, Git

Patent Applications and Disclosures

[3] **Akhil Arora**, Manoj Gupta, Neeta Pande, Sainyam Galhotra, Shourya Roy. *System for Identifying Root Causes of Churn for Churn Prediction Refinement.* USPTO Application Number: 15/132,767, Filed: 2016.

[2] **Akhil Arora**, Manoj Gupta, Shourya Roy. *Transforming a Knowledge Base into a Machine Readable Format for an Automated System.* USPTO Application Number: 14/887,096, Filed: 2015, Granted: 2018.

[1] **Akhil Arora**, Sainyam Galhotra, Srinivas Virinchi, Shourya Roy. *Methods and Systems for Identifying Target Users of Content.* USPTO Application Number: 14/628,070, Filed: 2015.

Publication Summary (cf. Google Scholar or List of Publications in Appendix for details)

30 peer-reviewed **papers** in top-tier Web and IR **(WWW, SIGIR, WSDM, ICWSM)**, NLP **(EMNLP, NAACL, LREC)**, and Data-centric **(SIGMOD, VLDB, EDBT)** venues. My **h-index is 10** and my publications have accrued a total of **652 citations.**

Honors and Awards

2023	**DAAD AInet Fellow** on Human-centered AI
2023	**Distinguished Reviewer Award**, CIKM
2023	**Distinguished Service Award** for contributions as a PhD representative, EPFL
2021	**Heidelberg Laureate Forum** participant (among 100 young researchers worldwide)
2018 – 2019	**EDIC Doctoral Fellowship**, EPFL
2018	**Most Reproducible Paper Award**, SIGMOD
2013	**Won** the **Adobe Data Mining Competition** held at IIT Madras
2012	Awarded the **Best Hack prize** in **Yahoo! HackU!** held at IIT Kanpur
2012	Stood **Second** in the 10th **ImageCLEF Machine Learning Challenge**
2012 – 2022	**Travel Grants Awarded**: VLDB' 22, SIGIR' 22, WSDM' 22, EDBT' 19, CLEF' 12
2019 – Present	**Raised** ∼**150K Euros** in the form of external funding for my research

Mentoring and Advising Experience

Extensive advising experience: In the past 10 years of my research career, I have mentored **over 25 students** for their research leading to **8 publications** to date. My mentees have been quite successful: as **PhD students/Lecturers** in reputed academic institutions (e.g. UC Berkeley, CMU, NUS) and **Engineers/Scientists** in reputed companies (Google, Meta, Expedia, UBS)

Invited Talks

2023 *AI-Assisted Knowledge Navigation*
- **Université Claude Bernard Lyon 1 and CNRS LIRIS** (Dec 2023)

2023 *Orphan Articles: The Dark Matter of Wikipedia*
- **Max Planck Institute for Software Systems (MPI-SWS)** (Dec 2023)
- **Expedia Group** (Jun 2023)

2022–2023 *Wikipedia Reader Navigation: When Synthetic Data is Enough*
- **Wikimedia Research Showcase** (Oct 2023)
- **Google India, Bangalore, India** (Apr 2022)

2022 *The Multiple Facets of Human Navigation on the Web*
- **University of Sydney, Sydney, Australia** (Sep 2022)
- **RMIT University, Melbourne, Australia** (Sep 2022)
- **Indian Institute of Technology (IIT), Delhi, India** (Jun 2022)

2020 *Low-rank Subspaces for Entity Linking without Annotated Data*
- **Utah Data Science Seminar** (Nov 2020)
- **Swiss Machine Learning Day (SMLD)** (Nov 2019)

2017 *Debunking the Myths of Influence Maximization*
- **American Express AI Labs** (Oct 2017)
- **University of Michigan, Ann Arbor, USA** (May 2017)
- **Indian Institute of Technology (IIT), Kanpur, India** (Feb 2017)

2015–2016 *Holistic Influence Maximization: Scalability and Efficiency with Opinion-Aware Models*
- **University of California, Santa Barbara, USA** (Jun 2016)
- **Facebook Inc., Menlo Park, USA** (Jun 2016)
- **Palo Alto Research Centre (PARC), USA** (Jun 2016)
- **Indian Institute of Technology (IIT), Kanpur, India** (Mar 2015)

2014–2015 *Mining Statistically Significant Connected Subgraphs in Vertex Labeled Graphs*
- **Palo Alto Research Centre (PARC), USA** (Sep 2015)
- **Xerox Research Centre Europe, Grenoble, France** (Oct 2014)
- **Xerox Research Centre India, Bengaluru, India** (Jun 2014)

Professional Service

2021 – 2024 **Steering Committee** – GRADES-NDA Workshop (Co-located with **SIGMOD**)

2016 – 2020 **PC Co-Chair** – GRADES-NDA Workshop (Co-located with **SIGMOD**)

2017 – Present **PC Member** – WSDM (2024, 2023, 2022, 2021), WWW (2024, 2022), ACL (2024, 2023), EMNLP (2023, 2022), EACL 2023, KDD (2024, 2023, 2022, 2021), AAAI (2022, 2021), CIKM (2023, 2022, 2021), EDBT (2024, 2021, 2020), Wikimedia Research Fund (2024, 2023), Wiki Workshop 2023, ICDE (2020 Demo), SIGMOD (2018 Demo), DASFAA (2017–2020), ISWC (2018–2020 Posters), AIMLSystems 2023

2017 – Present **Reviewer** – EMNLP (2021, 2020), SIGMOD 2019, Distinguished Reviewer Board: ACM Trans. on the Web (TWEB), ACM Trans. on Database Systems (TODS), VLDB Journal, IEEE Trans. on Knowledge and Data Engineering (TKDE), IEEE Trans. on Networks (ToN), ACM Trans. on Knowledge Discovery from Data (TKDD)

2013–2020 **External Reviewer** – VLDB (2016–2020), KDD (2015–2020), WWW (2019, 2017), ICDM (2019), ICDE (2021, 2020, 2018), SDM (2016, 2015), CIKM (2018, 2015, 2014)

2020 – Present **Ambassador**, EPFL IC Doctoral school (EDIC)

2022 – Present **PhD student representative**, EPFL IC Doctoral school (EDIC)

2023 – Present **Panelist** for the **Black in AI's** Emerging Leaders In AI Program

2018 – 2020 **Vice President**, EPFL IC PhD Association (EPIC)

2015 – 2016 **Organizing committee** member, XRCI Open 2015, Bangalore, India

2014 **Programming challenge Co-chair**, CoDS-COMAD

2012 – 2013 **Overall recruitment coordinator**, Students' placement office, IIT Kanpur

2012 – 2013 **Graduate student representative** for all computer science students, IIT Kanpur

Akhil Arora

+41 779876871
akhil.arora@epfl.ch
dlab.epfl.ch/people/aarora/

List of Publications

Refereed Publications

Indicators: *co-first authorship; †student I mentored.

Articles under review/preparation

[19] **Akhil Arora**, Robert West. *Large Language Models for Navigating the Web: The Case of Targeted Navigation on Wikipedia*. 2023.

[18] **Akhil Arora**, Martin Gerlach, Sayan Ranu, Robert West. *Link Recommendations for De-orphanizing Wikipedia Articles*. 2023.

[17] Marko Čuljak,*† **Akhil Arora**,* Andreas Spitz, Robert West, Karin Verspoor. *A Unified and Generic Benchmark for Entity Linking*. 2023.

[16] Veniamin Veselovsky,† Manoel Horta Ribeiro, **Akhil Arora**, Martin Josifoski, Ashton Anderson, Robert West. *Generating Faithful Synthetic Data with Large Language Models: A Case Study in Computational Social Science*. 2023.

Articles published at Conferences and Journals

[15] **Akhil Arora**, Robert West, Martin Gerlach. *Orphan Articles: The Dark Matter of Wikipedia*. In: Proc. of AAAI International Conference on Web and Social Media **(ICWSM)**, 2024. De-orphanization **tool:** https://linkrec.toolforge.org/

Top 5% of accepted papers

[14] Tiziano Piccardi, Martin Gerlach, **Akhil Arora**, Robert West *A Large-Scale Characterization of How Readers Browse Wikipedia*. ACM Trans. on the Web **(TWEB)**, 2023.

[13] **Akhil Arora**, Martin Gerlach, Tiziano Piccardi, Alberto García-Durán, Robert West. *Wikipedia Reader Navigation: When Synthetic Data is Enough*. In: Proc. of ACM International Conference on Web Search and Data Mining **(WSDM)**, 2022. **(Oral)**

Top 7% of accepted papers

[12] Manuel Leone,*†Stefano Huber,*†**Akhil Arora**,* Alberto García-Durán,* Robert West. *A Critical Re-evaluation of Neural Methods for Entity Alignment*. In: Proc. of International Conference on Very Large Data Bases **(PVLDB)**, 2022. **(Oral)**

[11] Vuk Vuković,†**Akhil Arora**, Huan-Cheng Chang,†Andreas Spitz, Robert West. *Quote Erat Demonstrandum: A Web Interface for Exploring the Quotebank Corpus*. In: Proc. of ACM International Conference on Research and Development in Information Retrieval **(SIGIR)** Demonstrations Track, 2022. https://quotebank.dlab.tools/

[10] Alberto García-Durán, **Akhil Arora**, Robert West. *Efficient Entity Candidate Generation for Low-Resource Languages*. In: Proc. of Language Resources and Evaluation Conference **(LREC)**, 2022. **(Oral)**

[9] **Akhil Arora**, Alberto García-Durán, Robert West. *Low-Rank Subspaces for Unsupervised Entity Linking*. In: Proc. of Conference on Empirical Methods in Natural Language Processing **(EMNLP)**, 2021.

[8] Jithin Vachery,†**Akhil Arora**, Sayan Ranu, Arnab Bhattacharya. *RAQ: Relationship Aware Graph Querying in Large Networks*. In: Proc. of The Web Conference **(WWW)**, 2019. **(Oral)**

[7] **Akhil Arora**, Sakshi Sinha,†Piyush Kumar,†Arnab Bhattacharya. *HDIndex: Pushing the Scalability-Accuracy Boundary for Approximate kNN Search in High Dimensional Spaces*. In: Proc. of Int. Conf. on Very Large Data Bases **(PVLDB)**, 2018. **(Oral)**

[6] **Akhil Arora**,* Sainyam Galhotra,* Sayan Ranu. *Debunking the Myths of Influence Maximization: An In-Depth Benchmarking Study*. In: Proc. of ACM International Conference on Management of Data **(SIGMOD)**, 2017. **(Oral)**

SIGMOD Most Reproducible Paper Award; Top 2% of accepted papers
SIGMOD 2017's Most Influential Paper #8

[5] Sainyam Galhotra,* **Akhil Arora**,* Shourya Roy. *Holistic Influence Maximization: Combining Scalability and Efficiency with Opinion-Aware Models*. In: Proc. of ACM International Conference on Management of Data **(SIGMOD)**, 2016. **(Oral)**

[4] Satyajit Bhadange,†**Akhil Arora**, Arnab Bhattacharya. *GARUDA: A System for Large-Scale Mining of Statistically Significant Subgraphs*. In: Proc. of International Conference on Very Large Data Bases **(PVLDB)** Demonstrations Track, 2016.

[3] Sainyam Galhotra,* **Akhil Arora**,* Srinivas Virinchi,†Shourya Roy. *ASIM: A Scalable Algorithm for Influence Maximization under the Independent Cascade Model*. In: Proc. of The Web Conference **(WWW)** Poster Track, 2015.

[2] Deepali Semwal, Sonal Patil, Sainyam Galhotra, **Akhil Arora**, Narayanan Unny. *STAR: Real-time Spatio-Temporal Analysis and Prediction of Traffic Insights using Social Media*. In: Proc. of ACM Joint International Conference on Data Science and Management of Data **(CoDS-COMAD)**, 2015.

[1] **Akhil Arora**, Mayank Sachan, Arnab Bhattacharya. *Mining Statistically Significant Connected Subgraphs in Vertex Labeled Graphs*. In: Proc. of ACM International Conference on Management of Data **(SIGMOD)**, 2014. **(Oral)**

Tutorials

[3] **Akhil Arora**,* Sainyam Galhotra,* Sayan Ranu. *Navigating the Maze of Influence Maximization Algorithms*. In: Proc. of IEEE International Conference on Data Science and Advanced Analytics **(DSAA)**, 2019.

[2] **Akhil Arora**,* Sainyam Galhotra,* Sayan Ranu. *Influence Maximization Revisited: The State of the Art and the Gaps that Remain*. In: Proc. of Extending Database Technology Conference (**EDBT**), 2019.

[1] **Akhil Arora**,* Sainyam Galhotra,* Sayan Ranu. *Influence Maximization Revisited: The State of the Art and the Gaps that Remain*, In: Proc. of ACM Joint International Conference on Data Science and Management of Data **(CoDS-COMAD)**, 2018.

Workshops and Symposiums

[13] **Akhil Arora**, Martin Gerlach, Robert West. *Orphan Articles: The Dark Matter of Wikipedia*. Int. Conference on Computational Social Science **(IC2S2)**, 2023. **(Oral)**

[12] Veniamin Veselovsky,†**Akhil Arora**, Tiziano Piccardi, Ashton Anderson, Robert West. *The Webonization of Wikipedia: Characterizing Wikipedia Linking Across the Web*. International Conference on Computational Social Science **(IC2S2)**, 2023. **(Oral)**

[11] Tiziano Piccardi, Martin Gerlach, **Akhil Arora**, Robert West. *A Large-Scale Characterization of How Readers Browse Wikipedia*. International Conference on Computational Social Science **(IC2S2)**, 2023.

[10] Veniamin Veselovsky,†**Akhil Arora**, Tiziano Piccardi, Ashton Anderson, Robert West. *The Webonization of Wikipedia: Characterizing Wikipedia Linking Across the Web*. **Wiki** Workshop, 2023. **(Oral)**

[9] Marko Čuljak,†Andreas Spitz, Robert West, **Akhil Arora**. *Strong Heuristics for Named Entity Linking*. In: Proc. of the North American Chapter of the Association for Computational Linguistics **(NAACL)** Student Research Workshop, 2022.

[8] **Akhil Arora**, Martin Gerlach, Tiziano Piccardi, Alberto García-Durán, Robert West. *Wikipedia Reader Navigation: When Synthetic Data is Enough.* Applied Machine Learning Days **(AMLD)**, 2022. **(Oral)**

[7] **Akhil Arora**. *Low-rank Subspaces for Entity Linking.* Youth in High Dimensions'22.

[6] **Akhil Arora**, Alberto García-Durán, Robert West. *Entity Linking via Low-rank Subspaces.* Swiss Machine Learning Day **(SMLD)**, 2019. **(Oral)**

[5] **Akhil Arora**,* Sainyam Galhotra,* Sayan Ranu. *Debunking the Myths of Influence Maximization.* North East Database Day **(NEDB)**, 2017. **(Oral)**

[4] Sainyam Galhotra,* **Akhil Arora**,* Shourya Roy. *Holistic Influence Maximization.* North East Database Day **(NEDB)**, 2016.

[3] **Akhil Arora**, Sumanth Naropanth. *Android Kernel and OS Security Assessment with Iron Crow.* **Black Hat** Europe, 2014. **(Oral)**

[2] Shashwat Mishra, Tejas Gandhi, **Akhil Arora**, Arnab Bhattacharya. *Efficient Edit Distance based String Similarity Search using Deletion Neighborhoods.* In: Proceedings of the Joint **EDBT/ICDT** Workshops, 2013. **(Oral)**

[1] **Akhil Arora**, Ankit Gupta, Nitesh Bagmar, Shashwat Mishra, Arnab Bhattacharya. *A Plant Identification System using Shape and Morphological Features on Segmented Leaflets.* In: Proc. of **CLEF** Workshops, 2012. **(Oral)**

References

Prof. Robert West
Assistant Professor
School of Computer and Communication Sciences
EPFL, Switzerland
✉ robert.west@epfl.ch

Dr. Martin Gerlach
Senior Research Scientist
Wikimedia Foundation
Berlin, Germany
✉ mgerlach@wikimedia.org

Prof. Karin Verspoor
Professor and Dean
School of Computing Technologies
RMIT University, Australia
✉ karin.verspoor@rmit.edu.au

Prof. Sayan Ranu
Associate Professor
Department of Computer Science
IIT Delhi, India
✉ sayanranu@cse.iitd.ac.in

Dr. Manish Gupta
Director and Head
Google Research India
Bangalore, India
✉ manishgupt@google.com

Prof. Andreas Spitz
Assistant Professor
Department of Computer Science
University of Konstanz, Germany
✉ andreas.spitz@uni-konstanz.de