Debjyoti Paul

December, 2022

CONTACT
dr.debjyotipaul@gmail.com
/in/debjyotipaul385/
www.debjyotipaul.in
www.github.com/debjyoti385

RESEARCH INTERESTS
Database Optimization, Large-Scale Data Analytics, Representation Learning, Spatio-Temporal Data, Natural Language Processing, Speech Technologies, Representation Learning.

EDUCATION

[2015-2020] Ph.D., University of Utah, School of Computing, Salt Lake City, Utah, USA
Major : Computer Science & Engineering
Advisor : *Prof. Feifei Li*
Thesis : *Data-Driven Spatio-Temporal Analysis for Multimode Data*
GPA : *3.96/4.0*

[2011-2013] Masters of Technology, Indian Institute of Technology Kanpur, Kanpur, India
Major : *Computer Science & Engineering*
Advisor : *Late Prof. Sanjeev K Aggarwal*
Thesis : *Multi-constraint Job scheduling in Grid Computing*
GPA : *8.67/10.0 (Rank: 3)*

[2007-2011] Bachelors of Technology, West Bengal University of Technology, Kolkata, India
Major : *Computer Science & Engineering*
College : *Institute of Enginering and Management*
CGPA : *8.93/10 (Rank: < 10)*

PROFESSIONAL EXPERIENCE

[March 2020 - *Present*] Meta, Inc., Menlo Park, United States
[Feb 2022 - *Present*] **Senior Research Scientist**
[Mar 2020 - Feb 2022] **Research Scientist**
Responsible for creating, improving, and deploying state-of-the-art automated speech recognition, Inverse Text Normalization and other companion multi-modal models.

[Aug 2015 - Jan 2020] University of Utah., Research Assistant, Salt Lake City, United States
Built a large-scale and efficient spatio-temporal data analytics framework.
Temporal data summarization: Historical or back-in-time bursty event detection.
Cloud database optimization and workload-aware database tuning with ML/AI.
Spatial data enrichment: Semantic representation for region of interest.

[Summer 2019] Alibaba Group., AI Research Intern, Sunnyvale, United States
Cloud database optimization and workload-aware database tuning with ML/AI.

[Summer 2018] Facebook, Inc., Research Scientist Intern, Seattle, United States
Improving metrics for Facebook's Search Ranking via rated label inferences with Machine Learning.

[Summer 2017] Amazon AI Research, Research Intern, New York, United States
Hyperparameter optimization techniques on MXNET.

[2013-2015] Flipkart, Data Engineer, Bangalore, India
Served as data engineer at data platform team *flipkart.com*, India's biggest e-commerce company.
Facilitated scalable environment for Big data analytics with processing pipelines for batch and stream.

THESES

[2020] Ph.D. Thesis, Advisor: Dr. Feifei Li
Data-Driven Spatio-Temporal Analysis for Multimode Data. [PDF]|[PPT]
The objective of this thesis is to build a framework that enables a data-driven approach for spatio-temporal analysis for multimode data with AI models and facilitate workload-aware support for effective large-scale processing.

[2013] M.Tech Thesis, Advisor: Late Dr. Sanjeev Kumar Aggarwal
Multi-constraint Job scheduling in Grid Computing. [PDF]|[PPT]
The objective is to efficiently schedule Jobs on Grid Computing to achieve maximum utilization of resources with energy efficient approach. Modeling real world computing & storage grid on Multi-objective Evolutionary Algorithm satisfying hard constraints on jobs constraints, resources constraints, and soft constraints on cost, and energy consumption.

Publications

+ Language Agnostic Data-Driven Inverse Text Normalization, Szu-Jui Chen, **Debjyoti Paul**, Yutong Pang, Peng Su, Xuedong Zhang. *ICASSP 2023. (In Submission).*

+ Improving Data Driven Inverse Text Normalization using Data Augmentation and Machine Translation, **Debjyoti Paul**, Yutong Pang, Sray Chen, Xuedong Zhang, *Interspeech 2022 Show and Tell*, Incheon, Korea, September 2022.

+ Improving Data Driven Inverse Text Normalization using Data Augmentation, Laxmi Pandey, **Debjyoti Paul**, Pooja Chitkara, Yutong Pang, Xuedong Zhang et al. CS arXiv:2207.09674 (2022).

+ Database Workload Characterization with Query Plan Encoders, **Debjyoti Paul**, Jie Cao, Feifei Li, Vivek Srikumar, *48th International Conference on Very Large Data Bases (VLDB 2022)*, Sydney, Australia, 2022. *DOI: 10.14778/3503585.3503600*

+ Semantic Embedding for Regions of Interest, **Debjyoti Paul**, Feifei Li, Jeff M. Phillips, *The International Journal on Very Large Data Bases (VLDBJ 2021). DOI:* 10.1007/s00778-020-00647-0.

+ Bursty Event Detection Throughout Histories, **Debjyoti Paul**, Yanqing Peng, Feifei Li, *35th IEEE International Conference on Data Engineering (ICDE 2019)*, Macau, China, 2019. *DOI:* 10.1109/ICDE.2019.00124

+ Compass: Spatio Temporal Sentiment Analysis of US Election, **Debjyoti Paul**, Feifei Li, Murali Krishna Teja, Xin Yu, Richie Frost, *What twitter says!*, *23rd SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD)*, Aug 13-17, 2017, Halifax, Canada. *DOI:* 10.1145/3097983.3098053.

+ AI Pro: Data Processing Framework for AI Models, **Debjyoti Paul**, Richie Frost, Feifei Li, *35th IEEE International Conference on Data Engineering (ICDE)*, Macau, China, 2019. *DOI:* 10.1109/ICDE.2019.00219.

+ Geotagged US Tweets as Predictors of County-Level Health Outcomes, 2015–2016, Quynh C. Nguyen, Matt McCullough, Hsien-wen Meng, **Debjyoti Paul**, Dapeng Li, *American Journal of Public Health (AJPH), September, 2017*, *DOI:* 10.2105/AJPH.2017.303993.

+ Social media indicators of the food environment and state health outcomes, Quynh C. Nguyen, Hsien-wen Meng, Dapeng. Li, Matt McCullough, **Debjyoti Paul**, Kanokvimankul. P, Nguyen. T, Li. Feifei, *American Public Health Association (APHA), 148, 120-128.*, 2017, *DOI:* 10.1016/j.puhe.2017.03.013.

+ Multi-objective Evolution based Dynamic Job Scheduler in Grid, **Debjyoti Paul**, Sanjeev K. Aggarwal, *The 8th International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS)*, IEEE, July 2nd - 4th, 2014, Birmingham, UK. *DOI:* 10.1109/CISIS.2014.50.

+ RCached-tree: An Index Structure for Efficiently Answering Popular Queries, Manash Pal, Arnab Bhattacharya, **Debjyoti Paul**, *ACM International Conference on Information and Knowledge Management (CIKM 2013)*, Oct. 27 - Nov. 1, 2013, San Francisco, CA, USA. *DOI:* 10.1145/2505515.2507817.

Position of Responsibilities

[2021] Program Committee Member for 47th International Conference on Very Large Data Bases (VLDB) 2021, Industrial Track.

[2021] Program Committee Member for 37th IEEE International Conference on Data Engineering (ICDE) 2021, Industrial Track.

[2018-2023] Reviewer: EDBT 2016, PVLDB (2019, 2021), IEEE TKDE (2018, 2019, 2020), IEEE ICDE 2021, ICASSP 2023.

[2015-2019] Teaching Mentor and Lecture: Natural Language Processing (CS6340), Data Mining (CS 6140), Database Systems (CS 6530) at the University of Utah. USA.

[2012-2013] Member and organizer of SIGDATA, a Special Interests Group on data management and mining at the IIT Kanpur, India.

[2011-2013] Teaching assistant: Software Engineering (CS455), Fundamentals of Computer Science (ESC101) at IIT Kanpur.

Select Projects

Workload Script
git:bit.ly/3hxClcb

This script automates installs of PostgreSQL, PostGIS 2.4, OLTP and OLAP benchmark on cloud instances. It also downloads and prepares respective benchmarks on cloud database instances with multiple database configurations. Additionally, it continuously uploads all the execution stats and metrics to a data collection server. This script is essential for database tuning research and training data generation.

OSM Benchmark
git:bit.ly/3PF1DS7

OSM is a database benchmark based on Open Street map dataset on Los Angeles, Salt Lake City and New York counties and cities containing spatial queries on multipolygons, points, and nearest neighbors.

Jackpine
git:bit.ly/3hxBKHt

JackPine is a geospatial database benchmark with updated support from PostGIS and updated shapefiles dataset.It contains spatial queries on multipolygons, lines, points and spatial joins among these combinations.

AI Pro
git:bit.ly/3YA7vQT
demo:bit.ly/3WlRLii

AI Pro is a framework for large scale Spatio-temporal Sentiment Analysis on social data. Example projects supported by Compass: Popularity of Political Parties for the Presidential Election 2016, Social Media indicators and social health outcomes. Read publications for details.

Event Aggregator
git:bit.ly/3Gb4qj7

This project finds events from news articles, categorize them and gathers all articles talking about same event to a set. Each set of articles is referred as Event Entity. Information extraction process is applied to extract more information related to it. The code can process large scale data. The code will be made public later.

QuestionAnswering
git:goo.gl/QGfW2y

A Natural Language Processing project focussed on closed domain Question Answering System. The project has Question Classifier, Sentence Similarity, Answer formulation and Coref resolution modules. The system has *Recall* of 63% and *F-score* of 43%.

Dartnews
git:bit.ly/3HMAbA2

A spatio-temporal news browsing application, with an interactive GIS interface, where news articles are organized by locality and topics such as crime, politics etc. Our topic and location prediction models has 87% accuracy, and 80% accuracy respectively.

QuakeAnalysis
git:bit.ly/3v2BXpb

This project is based on the seismic activity across world which widely varies in characteristic and patterns. We have found some distinguish patterns among the seismic activities and present them in an insightful manner. Check wiki of Exploration & Analysis

MusicAtlas
demo:bit.ly/3BLputP

This website is designed for music lovers to learn about music based on countries. This is a unique tool to explore and analyze the trend of music based on time frame, genre and artists. Almost all data from 19th century to till date.

RCached Tree
bit.ly/3YrOIXX

The objective is to speed up performance of point, range and kNN search queries for popular queries in databases. We came up with a variant of the R-Tree indexing structure with some features like caching at each node for popular queries.

Find more projects at https://www.debjyotipaul.in/

Hackathon Trophies

[2017]	HackTheU	HackTheU MLH Hackathon Winner (around 40 teams)
[2016]	EMC2 Code	Mars Challenge Hackathon Winner (around 45 teams)
[2015]	Goldman Sachs	Air Quality Hackathon 1st Runners-up (around 30 teams)
[2014]	InMobi	Freedom Hack Worldwide Hackathon 1st Runners-up (around 160 teams)
[2013]	Yahoo	Yahoo HackU 2013 Hackathon Winner (around 40 teams)

Achievements

[2015-2016] **Ph.D. Fellowship** at University of Utah.

[2013] Ranked 3rd out of 39 M.Tech students of CSE department in Indian of Institute Technology, Kanpur

[2012] Secured **All India Rank 228** out of 160,000+ (99.85 percentile) participants in Graduate Aptitude Test for Engineering (GATE) 2012, Computer Science.

[2011] Secured **All India Rank 223** out of 150,000+ participants in Graduate Aptitude Test for Engineering (GATE) 2011, Computer Science.

[2011] **All India Rank 7** (99.99 percentile) in Indian Space Research Organization (ISRO) Junior Research Fellow Exam.

[2010] 2nd Rank in Manual Robo-race competition and 2nd Rank in Robo-Olympics an innovative track based racing competition in Bits to Bytes Tech-fest, Bengal Institute of Technology.

References

1. **Feifei Li**
   Chief Database Scientist, Director, Alibaba Group, China.
   Ex-Professor, University of Utah, USA.
   Email: lifeifei@cs.utah.edu, ricosfeifei@gmail.com
2. **Jeff M Phillips**
   Associate Professor, School of Computing, University of Utah
   Email: jeffp@cs.utah.edu
3. **Quynh C. Nguyen**
   Associate Professor of Epidemiology and Biostatistics, University of Maryland, USA.
   Email: qtnguyen@umd.edu
4. **Vivek Srikumar**
   Associate Professor, School of Computing, University of Utah
   Email: svivek@cs.utah.edu