Debjyoti Paul

CONTACT	✉ debjyotipaul385@gmail.com 🏠 www.cs.utah.edu/~deb
	in /in/debjyotipaul385/ ⌂ www.github.com/debjyoti385

RESEARCH INTERESTS

Representation Learning, Language Translation, Sentiment Analysis, Natural Language Processing, Social Media Analysis, Spatio-Temporal Data, Public Health Analysis

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   : *3.98/4.0*

[2011-2013] Masters of Technology, Indian Institute of Technology Kanpur, Kanpur, India
Major   : *Computer Science & Engineering*
Advisor : *Late Professor 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)*

WORK EXPERIENCE

[March 2020 - *Current*] Facebook, Inc., Research Scientist, Menlo Park, United States
Working in the AI Speech team responsible for creating, improving, and deploying Automated Speech Recognition (ASR) models.

[Summer 2019] Alibaba Group., AI Research Intern, Sunnyvale, United States
Workload-aware database tuning with Machine Learning.

[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.
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
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

+ Semantic Embedding for Regions of Interest, **Debjyoti Paul**, Feifei Li, Jeff M. Phillips, *The International Journal on Very Large Data Bases (VLDBJ 2021).* [To Appear][PDF]

+ Database Workload Characterization with Query Plan Encoders, **Debjyoti Paul**, Jie Cao, Feifei Li, Vivek Srikumar, [In Review]

+ 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.

+ 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.

+ Lightweight Security Enhancement Protocol for Radio Frequency Identification (RFID), **Debjyoti Paul**, Sumana Basu, Punit Beriwal, *IEEE SPSITM International Conference 2011*, Kolkata, India.

POSITION OF
RESPONSIBILITIES

[**2021**] PC Member: ICDE 2021 Industrial Track.
[**2018-2021**] Reviewer: IEEE TKDE 2018, IEEE TKDE 2019, ICDE 2021, PVLDB 2021.
[**2016-2019**] External Reviewer: EDBT 2016, PVLDB 2019.
[**2015-2018**] Teaching Mentor: Natural Language Processing (CS6340), Data Mining (CS 6140) 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.

PROJECTS
(FOR MORE...
⌂/#PROJECTS)

Compass
estorm.org

Compass 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

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
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%.

Intelliad
goo.gl/UAnEdz

A Social Media driven Intelligent Ad-Targeting framework using Geo-profiling. The idea is to tag all geo-location enabled tweets(available publicly) with semantic categories (say sports, politics etc.) and their sentiment (positive, neutral, negative) using text mining. To enable serving of Ads, they also need to be tagged using same categories based on their content.

AirQuality
-https://goo.gl/F8twhn

Goldman Sachs Hackathon - This analyze the air quality of Utah state from real world dataset. Data warehousing and cubing to enable interactive visualization framework, which helped in determining pattern of Pollution and its reason.

Dartnews *goo.gl/wMOl4C*	A street news browsing application, with an interactive GIS interface. News is organized by locality and topic. The user can explore based on topics (eg crime, politics etc) and geolocation. We used Context Dependent Geoparsing, where we attempt to find out which location is relevant to the News. Across all the topics, we saw at least 87% accuracy of topic prediction, and at least 80% accuracy of location prediction.	
QuakeAnalysis *goo.gl/ovcWVT*	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 demo and wiki of Exploration & Analysis	
MusicAtlas	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.	
Real time discrimination of Speech and Music	Real-time speaker segmentation is required in many applications, such as speaker tracking in real- time news-video segmentation and classification, or real-time speaker adapted speech. After feature extraction, using Low Short Term Energy Ratio (LSTER) the input digital audio stream is classified into speech and music. Here we have used DSK6713 Digital Signal Processing Kit.	

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

[2015-2020] Ph.D. Fellowship at University of Utah.

[2012] **All India Rank 7** in Indian Space Research Organization (ISRO) Junior Research Fellow Exam.

[2011] Secured **All India Rank of 7** in Indian Space Research Organization (ISRO)'s Junior Research Fellow exam out of 200,000+ candidates.

[2012] Secured **All India Rank 223** out of 160,000+ participants and **All India Rank 228** out of 150,000+ participants in in Graduate Aptitude Test for Engineering (GATE) in 2011 and 2012 respectively in Computer Science.