Abhishek Roy

(413) 695-0872 • aroy@cs.umass.edu • cs.umass.edu/~aroy

Objective

Looking for full-time role in the areas of database systems and large-scale data analysis, starting from June 2018.

Interests

- Database Systems
- Distributed and Parallel Databases
- Large-Scale Data Analysis
- Genomic Data Analysis

Education

MS/PhD, University of Massachusetts Amherst 2011-May 2018
- PhD candidate in Computer Science
- Advisor: Prof. Yanlei Diao, Co-advisor: Prof. Prashant Shenoy
- GPA: 3.9/4

B.Tech, Indian Institute of Technology (IIT) Guwahati, India 2005-2009
- Bachelor of Technology in Computer Science and Engineering
- Thesis: Verifying Correctness of Dynamic Software Transactional Memory

Experience

Research Assistant (with Prof. Yanlei Diao), University of Massachusetts Amherst 2011-Present
- **Genomic Scalable Analysis with Low Latency (GESALL) Project:**
  - Built a **big data platform** for genomic data processing pipelines with an integrated solution to storage, data partitioning, runtime support, and high-quality results.
  - Introduced **Wrapper Technology** to parallelize existing genomic data analysis programs in their native forms, without having to rewrite them.
  - Collaborated with **New York Genome Center (NYGC)** to guarantee the quality of results.
  - Expert in measuring performance of Hadoop-based systems and multi-threaded programs.
  - Identified the key reasons behind super-linear and sub-linear speedups for different programs.
  - Reduced the turnaround time of deep analysis pipeline from two weeks to 1 day.
  - Code available at `https://gitlab.com/gesall`
- Previous work on association mining:
  - Identified causal relationship between genomic variations and genetic diseases.
  - Extended existing algorithms to support - a new interestingness metric, low support values, and proximity-aware mining.

Software Engineer, Strand Life Sciences, Bangalore, India 2009-2011
- Involved in the development of first version of Strand NGS, a software for next generation sequencing data analysis.
- Implemented algorithms and statistical tests to analyze large datasets.
- Contributed at every stage of product development life cycle.

Internships

NEC Laboratories, Cupertino, California Summer 2016
- Built a fast search engine to retrieve similar objects in high-dimensional feature space.
- Applied advanced database indexing techniques to outperform a parallel search system, based on Storm, by a factor of 10-100x.

Boston Children's Hospital, Boston, Massachusetts Summer 2012
- Worked in Genetic Diagnostic Laboratory to assemble and profile a deep analysis pipeline.

L3S Research Center, Hanover, Germany Summer 2008
- Developed a wavelet based approach to detect events from Flickr tags.

Publications

1. Extending Genome Data Parallel Toolkit to Support Multiple Data Models. *(in progress)*
2. **Abhishek Roy**, Yanlei Diao, Uday Evani, Avinash Abhyankar, Clinton Howarth, Rémi Le Priol, and Toby Bloom. Massively Parallel Processing of Whole Genome Sequence Data: An In-Depth Performance Study. SIGMOD 2017.
3. Yanlei Diao, **Abhishek Roy**, and Toby Bloom. Building Highly-Optimized, Low-Latency Pipelines for Genomic Data Analysis. CIDR 2015. *(main student author)*
4. **Abhishek Roy**, Yanlei Diao, Evan Mauceli, Yiping Shen, and Bai-Lin Wu. Massive Genomic Data Processing and Deep Analysis. VLDB 2012.
5. Ling Chen, and **Abhishek Roy**. Event Detection from Flickr Data through Wavelet-based Spatial Analysis. CIKM 2009. *(≈200 citations)*

Posters & Demonstrations

- "Massively Parallel Processing of Whole Genome Sequence Data: An In-Depth Performance Study" – SIGMOD 2017.
- "Building Highly-Optimized, Low-Latency Pipelines for Genomic Data Analysis" – NEDB 2015.
- "Massive Genomic Data Processing and Deep Analysis" – UMass Big Data Workshop 2013, NEDB 2013, VLDB 2012.

Teaching Experience

Teaching Assistant, Undergraduate Database Systems (CMPSCI 445) Spring 2014

- Designed and evaluated the course project.
- Delivered lectures on SQL and schema refinement.

Graduate Courses

Systems:		**Analysis:**	
Database Design and Implementation	A	Advanced Algorithms	A
Advanced Database Systems	A-	Artificial Intelligence	A-
Distributed Operating Systems	A	Probabilistic Graphical Models	A
Fast Algorithms and Parallel Processing	A	Mathematical Statistics I	A
Graduate Systems	A-	Machine Learning	Audit
Performance Evaluation	Audit	Integer Programming	Audit

Skills

Big Data Systems: YARN, HDFS, MapReduce, HBase, Spark
Databases: PostgreSQL, MySQL, RocksDB, MongoDB
Programming: Java, Python, C, C++, Go, R, MATLAB, AMPL, PHP, Assembly, Bash, Linux systems
Frameworks/Tools: CUDA, SciPy, Weka, Network programming, Spin Model Checker, Git, SVN
Genomic Toolkits: HTSJDK, PicardTools, GATK, Genome Browser
Performance Debugging: gdb, perf, sar, SystemTap, JDK profilers, Valgrind

Services

- External Reviewer for SIGMOD, 2016 and SIGMOD Record, 2015.
- Recruitment Team, Strand Life Sciences, 2010.