## GANESH RAMESH

**E-mail:** ganesh_ramesh@yahoo.com
**Web:** https://ganeshramesh.github.io

**Residence Address:**
5239 Fioli Loop,
San Ramon, CA 94582
**Phone:** (408) 409-3956


## SUMMARY


I am a serial Research Engineer, passionate about solving challenging problems requiring applications of Data Science, Artificial Intelligence and Machine Learning. I have over 15 years of experience in building and shipping products using large scale data as the backbone. I am very interested in exploring emerging applications of Generative AI both from an evaluation as well as an application standpoint.


## RESEARCH INTERESTS

Large Scale applications of AI/ML, Search and Information Retrieval, Recommender Systems, Web/Text Mining and NLP, Product Data Science.

## EDUCATION

**Postdoctoral Fellow, Database Systems Lab** **11/2003 – 10/2005**
**University of British Columbia**
Supervisors: Dr. Laks V.S. Lakshmanan, Dr. Raymond Ng

**Ph.D. Computer Science.** **10/2003**
**University at Albany, State University of New York(SUNY)**
Thesis: Data Mining Techniques for Frequent Itemsets: Construction and Analysis
Supervisor(s): Dr. William A. Maniatty, Dr. Daniel J. Rosenkrantz

**M.S. (by Research) in Computer Science and Engineering:** **06/1997**
**Indian Institute of Technology, Madras** (Now Chennai)
Thesis: Design and Implementation of Tutoring Systems: An Intermediate Control Approach

**B.E. Computer Science and Engineering** **05/1994**
**University of Madras**
Specialization: Parallel Processing

## RESEARCH AND INDUSTRIAL EXPERIENCE

**Ema Unlimited** Mountain View, CA
Research Engineer December 2023 - Present

Machine learning lead for Generative AI technologies. My primary role is to develop the fundamental generative AI technology that will power downstream applications developed by Ema. Currently involved in bench marking available Language Models and exploring strategies for LLM applications within the enterprise.

**Apple Inc.** Cupertino, CA
Principal Research Engineer, AIML March 2022 - November 2023
Principal Research Engineer, Maps Search August 2017 - February 2022

**Key Achievements - Siri Query Understanding - Tech Lead**

- An ML tool based on scientific principles that will provide insight about the quality of training data and help a user reason about model predictions. This tool was based on gradients and was extensively used for both spot fixing errors observed in prediction as well as for data cleaning.
- A robust address parser that was a fine tuned BERT model which was shipped for both EN and JP markets.

**Key achievements - Maps Search - Architect/ML Lead for Address Ranker**

1. Designed, implemented and productionized the end-to-end Machine Learning pipeline for building the address search ranker for various markets.
2. Replaced the manually built rule-based ranker which had become technically cumbersome to maintain.
3. Responsible for the project management including outbound communication as well as cross-team interaction for the project which spanned two years.
4. Shipped the ranker to 75% of address traffic by 2019 and the rest of the world by June 2020 serving hundreds of millions of queries worldwide across the entire Apple Device spectrum.
5. Led a cross team effort to enable the flow of probe data features which was one of the top 5 important features for the ranker model (Lifting the performance by 2 percentage points in observed online metrics)
6. Contributed to several architectural discussions and code/design reviews within the Maps team, including Query Tagging, Query Understanding, Feature engineering, Document Stores, Recall, Data Imputation, Machine Translation and Evaluation Metrics.

**Management and Administration:**

1. Mentored three junior team members in several aspects of Machine Learning, Search and Ranking. With my guidance and mentoring, I was able to transition the ownership of the ranker to them in Fall 2020 and focus on other projects.
2. Was an active mentor in the company mentorship program. Actively mentored 5 engineers across all functions within Apple.
3. Active organizer for the Apple ML summit. Was a part of the organization committee, the review committee and chaired the Knowledge Bases and Search Track for three consecutive years raising the bar for the quality of published and accepted work.

**SAP** Palo Alto, CA
Principal Data Scientist May 2016 - July 2017

1. Data exploration for intelligent procurement.
2. Enterprise search relevance and intelligent procurement agents.
3. Consulting Data Scientist for transforming personnel data and Data Privacy for Machine Learning.

Technologies/Languages: Java, Python, R, MySQL, MongoDB

**Yahoo! Inc. (Yahoo! GEMINI)** Sunnyvale, CA
Principal Research Engineer October 2015 - May 2016

1. Shipped two major improvements to sponsored search ad quality methods.

2. Using data driven analyses and metrics, worked with product managers, account managers (advertisers) and engineers to bridge the gap between advertiser expectations/happiness and the state of the ad serving system.

3. Designed data-driven metrics to engage with account managers to promote Yahoo sponsored search serving to advertisers.

Technologies/Languages: Java, Python, Pig, Hive, Hadoop

**Edmodo Inc.** San Mateo, CA
Principal Scientist/Engineer/Lead November 2013 - August 2015

1. Engineered a system to collect and curate educational content for recommendations. This system is a collection of general purpose and domain-specific code that curates and annotates resources extracted from the web for education.

2. Led a team that successfully shipped: a. A content recommendation service for Edmodo's assessment product. b. A content ecosystem that allows search and discovery of educational content complete with ratings and reviews of the products in the ecosystem. My contribution in code was in developing a significant part of the backend.

3. Designed and built a set of matching algorithms for educational objects. This allows for fast approximate match that powers many of our recommendations.

Technologies/Languages: Java, Python, Pig, Hive, Hadoop, MongoDB, MySQL

**LinkedIn Corporation (Product Data Sciences)** Mountain View, CA
Senior Research/Data Scientist April 2012 - November 2013

1. Designed and engineered the data pipeline for constructing session data for LinkedIn users based on their web and mobile page views.

2. Used features from session data to predict clickers (collaboration with my intern). The predictive model showed a 20% lift in predicting clickers when session based features were used in addition to existing features.

3. Engineered and shipped two location-based data products: a. Business traveler user segment which identified users who travel and b. Inferred location update which shipped as a part of the profile completion product.

4. Filed 4 patents in the areas of Inferred identity, recommendations and social reputation for professional social networks.

5. Engineered several data product prototypes: a. Gender inference for LinkedIn users. b. Recommending and Curating resources for learning skills. c. Aspirational, goal-oriented recommender system for LinkedIn users.

Technologies/Languages: Java, Python, Pig, Hive, R, Hadoop, MongoDB, MySQL

**Yahoo! Labs (Advertising Sciences)** Santa Clara, CA
Senior Research Engineer August 2009 - March 2012

1. Engineered parts of a state of the art supply forecasting system for large scale, complex multi-dimensional impression forecasting for Yahoo's Display Advertising platform.

2. Monitoring Tools and Metrics: Designed a comprehensive set of performance metrics for measuring stability, accuracy and consistency of forecasts for guaranteed display advertising contracts. Engineered the tools and data pipelines to compute these metrics.

3. Data Audits: Responsible for conducting several impactful data audits that validated the effectiveness of newly deployed serving algorithms and exposed several key issues leading to future improvements. Engineered the data pipeline and the tools to automate the performance of these audits.

4. Supply Landscape understanding: Applied pattern mining algorithms for identifying multidimensional regions that had a high sell through rate (areas of great demand for ads) and developed a methodology for navigating and presenting these regions. Also developed a method for identifying new regions with potential supply for those with high demand.

Technologies/Languages: Java, Python, Perl, Pig, Hive, Hadoop, MySQL

**Microsoft Corporation(Windows Live Safety Platform), Antispam** Redmond, WA
Software Development Engineer May 2007 - July 2009

1. **Mining Traceroute Data for Router Reputation:** Researched and engineered a scalable system that identifies and uses router statistics to filter spam e-mail. This system (in a conservative setting) helped reduce spam by 2%.

2. **Tools for Extracting Large Scale Network Data:** Engineered a multi-threaded generic network data collection tool for collecting network data (DNS, RDNS, Traceroute, Port scanners), significantly speeding up the throughput of data collection (upto 2 orders of magnitude in the best case) to baseline single threaded methods.

3. **Investigation of Natural Language Features for Classifying E-mail:** Studied the impact of generic natural language features extracted from graded mail to identify and classify spam e-mail.

4. **Tools for Automatic Extraction of Mail:** Engineered a tool for automatically extracting mails from various e-mail service providers (ESPs) and storing them into a structured relational database for spam statistics reporting.

Technologies/Languages: C#, SQL Server

**Microsoft Corporation(Live Search)** Redmond, WA
Software Development Engineer Jul 2006 - May 2007

1. **Mining and Monitoring Definitive Answers for Search Queries:** Definitive lists contain queries and their most relevant URLs in a cache for serving by the search engine, used as a failsafe mechanism. These lists are large, manually generated, multi-lingual, noisy and often ambiguous. I shipped language-insensitive methods based on click logs to identify potentially bad definitives. These techniques impacted upto 10% of the entries in many markets.

2. **Query Log Analysis for Information Vs Navigational Queries:** Performed a preliminary investigation of query and click logs to identify informational queries versus navigational queries - Queries for which people are seeking information spread across several URLs versus those that are satisfied by one single URL.

Technologies/Languages: C#, SQL Server

**Amazon.com, Customer Behaviour Group** Seattle, WA.
Analyst Oct 2005 - Jun 2006

1. **Optimizing Advertisement Content Delivery for Amazon Associates:** Designed, implemented and shipped a scalable machine learning model for recommending the best advertisements to be displayed on Amazon associate websites. Supported and further tuned the model to handle quality of recommendations and advanced phrasal features.

Technologies/Languages: Java, Perl, C++

**Database Systems Lab, University of British Columbia** Vancouver, B.C.
Postdoctoral Research Fellow Nov 2003 - Oct 2005

1. **Disclosure Risk Analysis of Anonymized Data:** We studied the problem of analyzing the risk in disclosing anonymized data to a third party. We considered the important case when the adversary or the third party has background knowledge about the data into our risk analysis. We proposed several practical methods using which a data owner can determine whether it is safe to release the data to the third party or not.

2. **Satisfiability of Tree Pattern Queries:** An important issue in evaluating XML queries is *satisfiability* i.e., given a query, can there exist a database on which the query has a non-empty answer (the database is expected to be consistent with respect to the schema, when one is present). We systematically studied satisfiability of tree pattern queries (which capture a useful fragment of XPath). We identified cases in which the problem can be solved in polynomial time and developed efficient algorithms for this purpose. We also identified several cases for which the problem is NP-complete. Our experience concludes through a systematic set of experiments that satisfiability check could effect substantial savings in query evaluation (Substantial savings for unsatisfiable queries and negligible overhead for satisfiable queries).

**Avaya Labs Research** Basking Ridge, New Jersey
Research Intern Aug 2001 - Nov 2001
**Text Mining for Multimedia Clustering:** Developed novel text mining techniques to learn patterns from a corpus of newswire articles. Developed a novel technique to filter commercial segments from closed caption text. Using the mined results from newswire articles, we developed efficient methods to improve the sparsity of closed caption text. These methods were successfully applied in building a hybrid clustering algorithm prototype for indexing and retrieving broadcast news data. The hybrid clustering algorithm significantly improved the quality of clusters by linking several related news articles that were previously unidentified.

**University at Albany, State University of New York** Albany, New York
Research Assistant Sep 2000 - May 2003

1. **Distribution-Based Synthetic Database Generation:** Analyzed feasible itemset distributions in databases and developed novel strategies for producing synthetic databases for itemset mining.

2. **Indexing and Data Access Methods for Data Mining** - Conducted an extensive performance analysis of the impact of access methods and data layouts on the behaviour of representative itemset mining algorithms. Our goal was to integrate various data mining algorithms with existing data management technology by decoupling the algorithms from the actual storage technique.

**Hewlett-Packard labs Research** Palo Alto, California
Research Intern May 1999 - Aug 1999
**Data Mining Applications:** Applied decision tree and association rule based techniques to analyze HP data. Proposed and developed an association rule mining workbench to study and benchmark the performance of association rule mining algorithms.

**General Electric Corporate Research and Development Labs** Schenectady, New York

Research Intern May 1998 - Aug 1998
**Customer Retention Models:** Applied decision tree and regression based methods to develop models for retaining GE customers. This was successfully used to identify a list of potentially bad customers (based on which further action was taken).

## TEACHING EXPERIENCE

**University of British Columbia** Vancouver, B.C.
**Instructor:** *Introduction to Relational Databases* - CPSC 304 Term: Summer 2004.

**University at Albany, SUNY** Albany, New York.
**Instructor:**
*Systems Programming* - CSI 402 Terms: Summer 2000, 2002, 2003.
*Introduction to Discrete Mathematics* - CSI 210 Term: Summer 2000.

**Teaching Assistant:**
*Algorithms and Data Structures* - CSI 403 Terms: Fall 1997, Fall 1999.
*Theory of Computation* - CSI 409 Terms: Spring 1998, Spring 1999.
*Principles of Programming Languages* - CSI 311 Terms: Spring 2002, Spring 2003.
*Programming the Hardware/Software Interface* - CSI 333 Term: Fall 1998.
*Database Systems* - CSI 508 Term: Spring 2000.

## PROFESSIONAL SERVICE

**Department of Computer Science, University at Albany, State University of New York**
Member of Faculty Search Committee 1999-2000
Member of Faculty Search Committee 2000-2001

**Organizing Committee Member**
Simulation Day, University at Albany, SUNY, October 2002.
Discrete Mathematics and Computer Science Day (DMCSD), University at Albany, SUNY, September 2002.
Network Design Day, University at Albany, SUNY, October 1998.

**Conference/Journal Referee**
**Program Committee Member:** WIDM 2006 (Web Information and Data Management), KDD 2007 (Industrial track), DASFAA 2008, ICDM 2011 - 2013, 2015 - 2019, CIKM 2018, DBSocial workshop 2013, IEEE Big Knowledge 2017.
**2009-2013:** Reviewer for IEEE TKDE, ACM TKDD and VLDB Journal
**2005:** External reviewer for ICDE, VLDB, SIGMOD. Reviewer for IEEE Transactions on Knowledge and Data Engineering, DAMI.
**2003, 2004:** External Reviewer for SIGMOD, PAKDD, SIGKDD, VLDB and ICDM Conferences. Reviewer for IEEE Transactions on Knowledge and Data Engineering and VLDB Journal.
**2002:** External Reviewer for SIAM Conference on Data Mining.

## AWARDS

**University Rank Holder:** Received a silver medal for obtaining the seventh highest overall GPA (out of over 1000 students) in my undergraduate program from the University of Madras.

## TALKS AND PRESENTATIONS

1. **To Do or Not To Do: The Dilemma of Disclosing Anonymized Data**, at University of Victoria, July 2008, at ACM SIGMOD Conference, June 2005 and at the Colloqium in Honor of Dr. Daniel J. Rosenkrantz, University at Albany, May 6, 2005.

2. **On Testing Satisfiability of Tree Pattern Queries**, VLDB Conference, September 2004.

3. **The Combinatorics of Frequent Itemset Mining**, Database Systems Lab, University of British Columbia, June 2003.

4. **Feasible Itemset Distributions in Data Mining: Theory and Application**, ACM PODS Conference, June 2003.

5. **Indexing and Data Access Methods for Data Mining**, ACM SIGMOD Workshop on Data Mining and Knowledge Discovery (DMKD), June 2002.

## PUBLICATIONS - Refereed Conferences and Workshops

1. Sean Chester, Bruce Kapron, Ganesh Ramesh, Gautam Srivastava, Alex Thomo and S. Venkatesh, **K-Anonymization of Social Networks by Vertex Addition**, in Proceedings II of $15^{th}$ East-European Conference on Advances in Databases and Information Systems (ADBIS), ADBIS 2011 Research Communications, 2011.

2. Shaofeng Bu, Laks V.S. Lakshmanan, Raymond Ng and Ganesh Ramesh, **Preservation of Patterns and Input-Output Privacy**. In IEEE International Conference on Data Engineering **ICDE**, Istanbul, Turkey, 2007.

3. Ganesh Ramesh, **Can Attackers Learn from Samples?** In $2^{nd}$ **VLDB** workshop on **Secure Data Management**, Trondheim, Norway, September 2005.

4. Ganesh Ramesh, Mohammed J. Zaki and William A. Maniatty, **Distribution-Based Synthetic Database Generation Techniques for Itemset Mining**. In $9^{th}$ International Database Engineering and Applications Symposium **IDEAS**, Montreal, Canada, July 2005.

5. Laks V.S. Lakshmanan, Raymond Ng and Ganesh Ramesh, **To Do or Not To Do; The Dilemma of Disclosing Anonymized Data**. In **ACM SIGMOD** Conference, Baltimore, Maryland, June 2005.

6. Laks V.S. Lakshmanan, Ganesh Ramesh, Wendy (Hui) Hwang and Jessica (Zheng) Zhao, **On Testing Satisfiability of Tree Pattern Queries**. In **VLDB** Conference, Toronto, Canada, August 2004.

7. Ganesh Ramesh, William A. Maniatty and Mohammed J. Zaki, **Feasible Itemset Distributions in Data Mining: Theory and Application**. In **ACM PODS** Conference, San Diego, California, June 2003.

8. Ganesh Ramesh, William A. Maniatty and Mohammed J. Zaki, **Indexing and Data Access Methods for Database Mining**. In ACM-SIGMOD Workshop on Data Mining and Knowledge Discovery **(DMKD)**, Madison, Wisconsin, June 2002.

9. Amit Bagga, Jianying Hu, Jialin Zhong and Ganesh Ramesh, **Multi-source Combined-Media Video Tracking for Summarization**. In $16^{th}$ International Conference on Pattern Recognition **(ICPR)**, Quebec City, Canada, August 2002.

10. Ganesh Ramesh and Amit Bagga, **A Text-based Method for Detection and Filtering of Commercial Segments in Broadcast News**. In International Conference on Language Resources and Evaluation **(LREC)**, Las Palmas, Canary Islands, Spain, May 2002.

11. Amit Bagga, Breck Baldwin and Ganesh Ramesh, **A Methodology for Cross-Document Coreference Over Degraded Data Sources**. In Recent Advances in Natural Language Processing **(RANLP)**, Tzigov Chark, Bulgaria, September 2001.

## PUBLICATIONS - Refereed Journals:

1. Sean Chester, Bruce M. Kapron, Ganesh Ramesh, Gautam Srivastava, Alex Thomo, S. Venkatesh, **Why Waldo befriended the dummy? k-Anonymization of social networks with pseudo-nodes**. In Social Network Analysis and Mining, Volume 3, Number 3, pages 381-399, 2013.
2. Laks V.S. Lakshmanan, Raymond Ng and Ganesh Ramesh, **On Disclosure Risk Analysis of Anonymized Itemsets in the Presence of Prior Knowledge**. In ACM Transactions on Knowledge Discovery in Data (TKDD), Volume 2, Number 3, 2008.

## PATENTS:

1. Techniques for Inferring a Location, United States 13/871,570, Filed April 2013.
2. Recommending Resources To Members Of A Social Network, United States 13/890,693, Filed May 2013.
3. Method and System to Provide Reputation Scores for a Social Network Member, United States 13/947,866, Filed August 2013.

## Programming Skills

**Languages:** Java, Python, C#, C, C++, SQL

**Employment Eligibility:** United States Citizen.

**REFERENCES**: Available upon request.