Balakrishnan (Murali) Narayanaswamy

San Jose, CA — muralibnarayan@gmail.com — (330) 524–4608
Webpage —— Google Scholar

Summary

- 10+ years at Amazon/AWS, Principal Machine Learning Scientist and manager since 2021. 15 years industry experience, 20+ years ML/AI experience.
- Straddling IC and manager roles, at Amazon I lead a team of 10 Applied Scientists, building learning- and optimization-driven systems at cloud scale.
- I've led science and and delivery for 10+ AWS 'keynote' or 're:Invent' launches.
- Each product/launch has hit 1000 - 100k+ customers and is the result of interesting applied research we published at top-tier venues - ICML, NeurIPS, KDD, SIGMOD and VLDB.
- 45+ Research Publications, 25+ patents - https://sites.google.com/site/muralibalki/research, KDD 2020 ADS Best Paper Award.
- Ph.D.(2011), M.S. (2005) Electrical and Computer Engineering, Carnegie Mellon University

Focus — Continual Learning in theory, practice and at scale

ML learning and AI offer a lot of promise in scaling out repetitive tasks. However, the world does not stand still - what was a good prediction, recommendation, action or forecast yesterday may not be a good one today. This leads to users being disappointed with ML and AI applications making stale predictions or repeating the same errors over and over. On the other hand, systems that evolve too jerkily are hard to use. I work on making ML algorithms robust to changing environments in the real world, for 1000-10k-100k or more customers over millions and billions of data points. Continual learning may require failure, but success at scale requires extracting learnings from failures automatically, so that our systems don't make the same mistake twice.

Selected Products with Associated Publications

- GenAI in Amazon SageMaker Unified Studio: Amazon Q Developer provides generative AI assistance within Amazon SageMaker Unified Studio for data discovery, data processing, SQL analytics, and machine learning workflows. Associated Publication ModelSwitching-ICLR workshop. Ongoing work, hopefully more to come at ICML.
- Q for Generative SQL: an out-of-the-box web-based SQL editor for Redshift, to simplify query authoring and increase your productivity by allowing you to express queries in natural language and receive SQL code recommendations. Associated Publications SQLens-NeurIPS, TailorSQL-AIDB show how to learn from agentic execution errors and adapt to ambiguity.
- Structured Knowledge Bases (Amazon Bedrock): Amazon Bedrock Knowledge Bases allows you to connect to structured data stores. Amazon Bedrock Knowledge Bases can convert user queries into language that is suitable for extracting data from supported structured data stores and generate appropriate responses. Associated Publications ODIN-arxiv

- AI-driven scaling & optimization in Redshift Serverless: scaling and optimization in cloud data warehousing. Amazon Redshift Serverless uses AI techniques to automatically scale with workload changes across all key dimensions—such as data volume changes, number of concurrent users, and query complexity to meet and maintain your price-performance targets. Associated Publications Sage-SIGMOD,Redset-VLDB, SageForecast-SoCC show how to adapt to changing query patterns and auto-scale through forecast-then-optimize.
- Automatic Workload Management in Redshift: manages query concurrency and memory allocation in Amazon Redshift, for billions of queries. Associated Publications Stage-VLDB, IconqSched-VLDB, AutoWLM-SIGMOD adapt execution based on query patterns through online retraining.
- Amazon Personalize: A fully managed Machine Learning (ML) service that uses your data to generate item recommendations for your users.. Sssociated Publications HRNN-KDD 2020 (Applied DS Track, Best Paper) shows how to design Personalization systems that automatically adapt to changing item popularity and user preferences by adding learned features.
- Amazon Forecast and (retail) demand forecasting: fully managed time-series forecasting service based on machine learning (ML) and built for business metrics analysis. Associated Publications MQRNN-NeurIPS workshop adapts buying and stocking to changes in seasons and popularity with RNNs.
- Amazon SageMaker RL: Managed Reinforcement Learning with Amazon SageMaker. Associated Publications ORL-arxiv, IROL-AISTATS
- AWS Lambda: VM placement in the cloud. Associated Publications Fireplace/MicroPlace-MLSys
- Amazon Redshift ML: makes it easy for data analysts and database developers to create, train, and apply machine learning models using familiar SQL commands in Amazon Redshift data warehouses.
- DevOpsGuru for RDS: ML-powered capability that is designed to empower developers and DevOps engineers to quickly detect, diagnose, and remediate a wide variety of database-related issues in Amazon RDS. Associated Publications CGATS-NeurIPS workshop, PANDA-CIDR, VISTA-SOCC make anomaly detection much less noisy, by adapting to customer usage patterns.
- GuardDuty RDS Protection: RDS Protection in Amazon GuardDuty analyzes and profiles RDS login activity for potential access threats to your Amazon Aurora and Amazon RDS databases. Associated Publications FITNESS-ICML, ORNE-NeurIPS, ODT-ICML,CP-UAI build up theoretical bounds and show how we deploy these algorithms at scale to secure cloud databases with billions of streamed security events.

Experience

Amazon (Retail / AWS AI / Learned Systems Group) 2015–Current
Principal Machine Learning Scientist who leads science teams and has built/launched/owned products in AWS AI, Bedrock, SageMaker Unified Studio, Redshift, and RDS.
UC San Diego - Machine Learning Group and MESL 2014–2015
Research Staff Member, IBM Research, Smarter Energy and HPC 2011–2013

45+ Research Publications, 25+ patents

2025

- Mintong Kang, Vinayshekhar Bannihatti Kumar, Shamik Roy, Abhishek Kumar, Sopan Khosla, Balakrishnan (Murali) Narayanaswamy, Rashmi Gangadharaiah, "FairGen: Controlling Sensitive Attributes for Fair Generations in Diffusion Models via Adaptive Latent Guidance," EMNLP 2025.
  Highlight: Proposes adaptive latent guidance to steer diffusion models away from sensitive attributes, improving fairness in generations.

- Yue Gong, Chuan Lei, Xiao Qin, Kapil Vaidya, Balakrishnan Narayanaswamy, Tim Kraska, "SQLens: An End-to-End Framework for Error Detection and Correction in Text-to-SQL," NeurIPS 2025.
  Highlight: End-to-end pipeline that detects and fixes NL2SQL errors to boost robustness in text-to-SQL systems.

- Kapil Vaidya, Jialin Ding, Sebastian Kosak, David Kernert, Chuan Lei, Xiao Qin, Abhinav Tripathy, Ramesh Balan, Balakrishnan Narayanaswamy, Tim Kraska, "TailorSQL: An NL2SQL System Tailored to Your Query Workload," arXiv:2505.23039.
  Highlight: Personalizes NL2SQL to a target workload for higher accuracy on in-domain schemas.

- Kapil Vaidya, Abishek Sankararaman, Jialin Ding, Chuan Lei, Xiao Qin, Balakrishnan Narayanaswamy, Tim Kraska, "ODIN: A NL2SQL Recommender to Handle Schema Ambiguity," arXiv:2505.19302.
  Highlight: Recommender that resolves schema ambiguities to suggest more faithful SQL translations.

- Ziniu Wu, Markos Markakis, Chunwei Liu, Peter Baile Chen, Balakrishnan Narayanaswamy, Tim Kraska, Samuel Madden, "Improving DBMS Scheduling Decisions with Fine-grained Performance Prediction on Concurrent Queries," VLDB 2025 (arXiv:2501.16256).
  Highlight: Fine-grained predictors for concurrent query performance to improve scheduler choices.

2024

- Vikramank Singh, Zhao Song, Balakrishnan (Murali) Narayanaswamy, Kapil E. Vaidya, Tim Kraska, "Vista: ML-based Database Performance Troubleshooting in Amazon RDS," ACM SoCC 2024.
  Highlight: Practical framework that localizes and explains DB performance issues using ML.

- Yanlei Diao, Dominik Horn, Andreas Kipf, Oleksandr Shchur, Ines Benito, Wenjian Dong, Davide Pagano, Pascal Pfiel, Vikram Nathan, Balakrishnan (Murali) Narayanaswamy, Tim Kraska, "Forecasting Algorithms for Intelligent Resource Scaling: An Experimental Analysis," ACM SoCC 2024.
  Highlight: Comparative study of forecasting methods that drive elastic scaling for cloud databases.

- Alexander van Renen, Dominik Horn, Pascal Pfeil, Kapil E. Vaidya, Wenjian Dong, Balakrishnan (Murali) Narayanaswamy, Zhengchun Liu, Gaurav Saxena, Andreas Kipf, Tim Kraska, "Why TPC is not enough: An analysis of the Amazon Redshift fleet," VLDB 2024.

Highlight: Fleet-scale study showing real-world workloads diverge from TPC benchmarks, motivating new evaluation.

- Sophia Huiwen Sun, Abishek Sankararaman, Balakrishnan (Murali) Narayanaswamy, "Online Adaptive Anomaly Thresholding with Confidence Sequences," ICML 2024.
  Highlight: Uses confidence sequences to adapt thresholds online for drift-robust anomaly detection.
- Vikram Nathan, Vikramank Singh, Zhengchun Liu, Mohammad Rahman, Andreas Kipf, Dominik Horn, Davide Pagano, Gaurav Saxena, Balakrishnan (Murali) Narayanaswamy, Tim Kraska, "Intelligent scaling in Amazon Redshift," SIGMOD 2024.
  Highlight: Learning-driven autoscaling for Redshift that reduces latency and cost under bursty workloads.
- Ziniu Wu, Ryan Marcus, Zhengchun Liu, Parimarjan Negi, Vikram Nathan, Pascal Pfeil, Gaurav Saxena, Mohammad Rahman, Balakrishnan (Murali) Narayanaswamy, Tim Kraska, "Stage: Query execution time prediction in Amazon Redshift," SIGMOD 2024.
  Highlight: Production time-prediction for Redshift queries to inform admission control and scheduling.
- Vikramank Y. Singh, Kapil Vaidya, Vinayshekhar Bannihatti Kumar, Sopan Khosla, Balakrishnan (Murali) Narayanaswamy, Rashmi Gangadharaiah, Tim Kraska, "Panda: Performance Debugging for Databases using LLM Agents," CIDR 2024.
  Highlight: LLM agents that reason over logs/metrics to diagnose DB slowdowns.

2023

- Abishek Sankararaman, Balakrishnan (Murali) Narayanaswamy, "Online Heavy-tailed Change-point Detection," UAI 2023.
  Highlight: Detects change-points with provable guarantees under heavy-tailed distributions.
- Abishek Sankararaman, Balakrishnan (Murali) Narayanaswamy, Online robust non-stationary estimation, NeurIPS 2023.
  Highlight: Online estimators that remain stable under regime shifts and heavy-tailed noise.
- Gaurav Saxena, Mohammad Rahman, Naresh Chainani, Chunbin Lin, George Caragea, Fahim Chowdhury, Ryan Marcus, Tim Kraska, Ippokratis Pandis, Balakrishnan (Murali) Narayanaswamy, Auto-WLM: machine learning enhanced workload management in Amazon Redshift, SIGMOD 2023.
  Highlight: Learns workload-aware resource allocation to improve QoS and throughput.

2022

- Kaustubh Sridhar, Vikramank Singh, Balakrishnan (Murali) Narayanaswamy, Abishek Sankararaman, "Predict-and-Critic: Accelerated End-to-End Predictive Control for Cloud Computing through Reinforcement Learning," under review.
  Highlight: Combines predictive models with critic signals to speed RL-based cloud control.
- Vikramank Singh, Abishek Sankararaman, Balakrishnan (Murali) Narayanaswamy, Zhao Song, "C-GATS: Conditional Generation of Anomalous Time Series," NeurIPS 2022 Workshop (SyntheticData4ML).
  Highlight: Conditional generative model that synthesizes controllable anomalies for evaluation.

- Abishek Sankararaman, Balakrishnan (Murali) Narayanaswamy, Vikramank Singh, Zhao Song, "FITNESS: (Fine Tune on New and Similar Samples) to detect anomalies in streams with drift and outliers," ICML 2022.
  Highlight: Transfer-style fine-tuning to stay accurate under drift and outliers in streams.

2021

- Bharathan Balaji, Christopher Kakovitch, Balakrishnan (Murali) Narayanaswamy, "MicroPlace: Placing Micro Virtual Machines with Hindsight Imitation," MLSys 2021.
  Highlight: Uses imitation learning with hindsight to place micro-VMs efficiently at scale.

2020

- Bharathan Balaji, Christopher Kakovitch, Balakrishnan (Murali) Narayanaswamy, "FirePlace: Placing FireCracker virtual machines with hindsight imitation," NeurIPS 2020 Workshop (ML for Systems).
  Highlight: Demonstrates learning-based placement for serverless microVMs.
- Yifei Ma*, Balakrishnan (Murali) Narayanaswamy*, Haibin Lin, Hao Ding, "Temporal-Contextual Recommendation in Real-Time," KDD 2020 (Applied DS Track, Best Paper).
  Highlight: Real-time recommender using temporal and contextual features; KDD ADS Best Paper.
- Rashmi Gangadharaiah, Balakrishnan (Murali) Narayanaswamy, "Recursive Template-based Frame Generation for Task-Oriented Dialog," ACL 2020 (Oral).
  Highlight: Template recursion to generate coherent task flows for dialog.

2019

- Yifei Ma, Yu-Xiang Wang, Balakrishnan (Murali) Narayanaswamy, "Imitation-Regularized Offline Learning," AISTATS 2019.
  Highlight: Introduces imitation regularization to stabilize and improve offline RL.
- Rashmi Gangadharaiah, Balakrishnan Narayanaswamy, "Joint Multiple Intent Detection and Slot Labeling for Goal-Oriented Dialog," NAACL-HLT 2019.
  Highlight: Joint model tying intents and slots for improved goal-oriented dialog understanding.
- Bharathan Balaji, Jordan Bell-Masterson, Enes Bilgin, Andreas Damianou, Pablo Moreno Garcia, Arpit Jain, Runfei Luo, Alvaro Maggiar, Balakrishnan Narayanaswamy, Chun Ye, "ORL: Reinforcement Learning Benchmarks for Online Stochastic Optimization Problems," arXiv:1911.10641.
  Highlight: Benchmark suite for online stochastic optimization problems in RL settings.

2017

- Rashmi Gangadharaiah, Balakrishnan Narayanaswamy, Charles Elkan, Achieving Fluency and Coherency in Task-oriented Dialog, NeurIPS 2017 Workshop.
  Highlight: Practical strategies to improve fluency and coherence in task bots.
- Ruofeng Wen, Kari Torkkola, Balakrishnan (Murali) Narayanaswamy, Dhruv Madeka, "A Multi-Horizon Quantile Recurrent Forecaster," NeurIPS 2017 (Time Series Workshop).
  Highlight: Quantile RNN that forecasts across multiple horizons for calibrated uncertainty.

2015 and earlier (selected)

- Bharathan Balaji, Chetan Verma, Balakrishnan Narayanaswamy, Yuvraj Agarwal, "Zodiac: Organizing Large Deployment of Sensors to Create Reusable Applications for HVAC Management," BuildSys 2015.
  Highlight: System for reusable, large-scale building sensor deployments.

- Xun Jiao, Abbas Rahimi, Balakrishnan Narayanaswamy, Hamed Fatemi, Jose Pineda de Gyvez, Rajesh Gupta, "A Scalable Model for Timing Error Prediction under Hardware and Workload Variations," NEWCAS 2015.
  Highlight: Scalable model for predicting timing errors under HW/workload variability.

- Pankaj Dayama, Balakrishnan Narayanaswamy, Dinesh Garg, Yadati Narahari, "Truthful Interval Cover Mechanisms for Crowdsourcing Applications," AAMAS 2015.
  Highlight: Incentive-compatible covering mechanisms for crowdsourced tasks.

- Bochao Shen, Balakrishnan Narayanaswamy, Ravi Sundaram, "SmartShift: Expanded Load Shifting Incentive Mechanism for Risk-averse Consumers," AAAI-15.
  Highlight: Demand-shifting mechanism accounting for consumer risk.

- Balakrishnan Narayanaswamy, Bharatan Balaji, Yuvraj Agarwal, Rajesh Gupta, "Data Driven Investigation of Faults in HVAC Systems with Model, Cluster and Compare (MCC)," BuildSys 2014.
  Highlight: Model/cluster/compare pipeline to isolate HVAC faults.

- Rizhen Zhang, Bharathan Balaji, Yan Zhang, Balakrishnan Narayanaswamy, Yuvraj Agarwal, "Demo Abstract—BuildingSherlock: Fault Management Framework for HVAC Systems," BuildSys 2014.
  Highlight: Fault-management framework for HVAC at scale.

- Muhammad Adnan, Balakrishnan Narayanaswamy, Rajesh Gupta, "Online Reservation and Deferral of EV Charging Tasks to Reduce Energy Use Variability in Smart Grids," SmartGridComm 2014.
  Highlight: Online policies to shift EV loads and cut variance.

- Shweta Jain, Balakrishnan Narayanaswamy, Y. Narahari, "A Multiarmed Bandit Incentive Mechanism for Crowdsourcing Demand Response in Smart Grids," AAAI-14.
  Highlight: Bandit-based demand-response incentives.

- Vijay Arya, Balakrishnan Narayanaswamy, "Loss Localisation in Smart Distribution Networks," COMSNETS 2014 (Best Paper Runner Up).
  Highlight: ML-based localization of losses in distribution networks.

- Farah Japar, Sathyajith Mathew, Balakrishnan Narayanaswamy, Chee Ming Lim, Jagabondhu Hazra, "Estimating the Wake Losses in Large Wind Farms: A Machine Learning Approach," IEEE ISGT 2014.
  Highlight: ML approach for wind-farm wake loss estimation.

- Neha Sengupta, Aloka S, Balakrishnan Narayanaswamy, Hamidah Ismail, Satyajith Mathew, "Time Series Data Mining for Demand Side Decision Support," ISGT-Asia 2013.
  Highlight: TS mining to aid demand-side decisions.

- Hamidah Ismail, Satyajith Mathew, Aloka S, Balakrishnan Narayanaswamy, Lim Chee Ming, Saiful Azmi Hussain, "Comparative Performances of Grid Integrated Solar Photovoltaic Sys-

tems Under the Tropical Environment," ISGT-Asia 2013.
Highlight: Field comparison of grid PV under tropical conditions.

- Chetan Yadati, Dilpreet Kaur, Balakrishnan Narayanaswamy, Avik Sarkar, Narahari Y, "Incentive Compatible Mechanisms for Power Cut Allocation in Smart Grids," IEEE CASE 2013.
Highlight: Incentive-compatible power cut allocation.
- V. K. Garg, T. S. Jayram, Balakrishnan Narayanaswamy, "Online Optimization with Dynamic Temporal Uncertainty: Incorporating Short Term Predictions for Renewable Integration in Intelligent Energy Systems," AAAI-13.
Highlight: Short-term predictions to improve online control.
- Shweta Jain, Balakrishnan Narayanaswamy, Y. Narahari, Saiful A. Hussain, Nyuk Yoong Voo, "Constrained Tâtonnement for Fast and Incentive Compatible Distributed Demand Management in Smart Grids," ACM e-Energy 2013.
Highlight: Fast, incentive-compatible distributed demand management.
- Jagabondhu Hazra, Kaushik Das, Ashok Pan Kumar, Balakrishnan Narayanaswamy, Deva P. Seetharam, Nis Jesperson, "Optimal Utilization of Power Transformers Through Virtual Sensing," ISGT-EU 2013.
Highlight: Virtual sensing to extend transformer utilization.
- Abishek S., Balakrishnan Narayanaswamy, "Congestion Control of Smart Distribution Grids using State Estimation," IEEE COMSNETS 2013, E6 Workshop.
Highlight: State-estimation-driven congestion control.
- Balakrishnan Narayanaswamy, T. S. Jayram, V. K. Garg, "Prediction based storage management in the smart grid," SmartGridComm 2012.
Highlight: Uses prediction to optimize storage operations.

Selected Patents out of 25+

- Assisted Database Anomaly Mitigation US 12,130,788
Highlight: Automates detection and guided mitigation of DB anomalies to restore performance quickly.
- Detecting idle periods at network endpoints for management actions at processing clusters for managed databases WO 2023097270 A1
Highlight: Finds safe "idle windows" so maintenance actions won't disrupt active workloads.
- Scaling query processing resources for efficient utilization and performance US Patent App. 17/547,831
Highlight: Dynamically scales query resources to balance utilization and latency for changing workloads.
- Dynamic compilation of machine learning models based on hardware configurations US 11,657,069
Highlight: Selects/compiles model binaries tailored to target hardware for speed and efficiency.
- Integrating query optimization with machine learning model prediction US 11,636,124
Highlight: Uses ML predictions inside the query planner to improve cost estimates and plans.
- Automatically refreshing materialized views according to performance benefit US 11,609,910
Highlight: Prioritizes MV refreshes by expected end-to-end performance impact.

- Dynamic generation of match rules for rewriting queries to use materialized views US 11,550,787
  Highlight: Auto-derives rewrite rules so queries transparently benefit from materialized views.
- Goal-oriented dialog systems and methods US 10,963,819
  Highlight: Dialog framework with goal/slot tracking for robust task completion.
- Task-oriented dialog systems using combined supervised and reinforcement learning US 10,860,629
  Highlight: Blends supervised and RL signals so assistants learn effective conversation strategies.
- Demand forecasting via direct quantile loss optimization US 10,783,442
  Highlight: Trains forecasters directly on quantile loss for calibrated predictions.
- Non-technical loss detection and localization US 9,536,198
  Highlight: Detects and localizes power-theft/non-technical losses in electric grids.
- Power factor US 9,343,904
  Highlight: Techniques to estimate and improve power factor for grid efficiency.
- Condition-based management of power transformers US 9,312,695
  Highlight: Uses condition data to optimize transformer maintenance and lifespan.
- Enforcing fine-grained demand management in smart grids US 9,188,967
  Highlight: Enforces granular demand-response policies across grid participants.
- Estimating loss rates of links in smart grids US 9,048,664
  Highlight: Estimates distribution-network link losses to target remediation.
- Estimating loss rates of links in smart grids US 9,043,174
  Highlight: Companion methods that improve accuracy of loss-rate estimation.