Cagri Ozcaglar

Sr. Machine Learning Engineer
Facebook (Meta)
Menlo Park, CA

cagriozcaglar@gmail.com
http://www.cs.rpi.edu/~ozcagc2
https://github.com/cagriozcaglar

Education

- **Ph.D., Computer Science** 2008-2012
  Rensselaer Polytechnic Institute, Troy, NY
  Thesis: Algorithmic Data Fusion Methods for Tuberculosis
- **M.S., Computer Science** 2006-2008
  Rensselaer Polytechnic Institute, Troy, NY
  Thesis: Classification of Email Messages into Topics Using Latent Dirichlet Allocation
- **B.S., Computer Science** 2002-2006
  Bilkent University, Ankara, Turkey

Skills

- **Languages:** Java, Python, Scala, C++, R, MATLAB, Shell Scripting, Groovy, PHP, Perl.
- **Modeling Software:** Tensorflow, Keras, PyTorch, Python (scikit-learn), R, MATLAB, Amazon Machine Learning, ProML.
- **Big Data Processing:** Spark, Hadoop, Apache Pig, AWS.
- **Database Management:** SQL, PostgreSQL, MySQL, Oracle 10g/11g.
- **Tools and Applications:** Latex, Eclipse, IntelliJ, Git, SVN.

Research Interests

Machine Learning Applications and Theory, Search and Recommendation Systems, Retrieval and Ranking Relevance, User Intent Modeling, Social Network Analysis, Machine Learning, Data Mining, Data Fusion, Multiway Analysis.

Experience

- **Sr. Machine Learning Engineer, Facebook (Meta)** January 2022 - Present
  Sr. Machine Learning Engineer Menlo Park, CA
  - Applied Machine Learning.
- **Sr. Machine Learning Engineer, LinkedIn** 2016 - October 2021
  Sunnyvale, CA
  Sr. Machine Learning Engineer 2018 - October 2021
  Teams: Down-Funnel Optimization, Careers Relevance
  - Team Tech lead as of April'21: Created and led a team charter for two-sided Recruiter & Jobs Marketplace optimization: 1) Semi-supervised learning using graded-strength labels, 2) Causal inference for uplift estimation, 3) Global inference using LP for multi-objective optimization (relevance, diversity, facepalm), 4) Multi-task learning for Recruiter & Jobs Search and Recommendation system pairs.
  - Mentored 4 new hires in their first 90-day ramp-up, and multiple internal hires to my team / org.
  - Won Data All Star Team Award at LinkedIn for 2020.
  - Led multiple efforts for improving key metrics and AI productivity in Job Recommendation system (JYMBII: Jobs You May Be Interested In) ranking:
    - Led and implemented Job Recommendation system ranking component migration from GLMix model to deep neural network models using Tensorflow. Contributed to all components including data preprocessing and preparation, modeling, serving and infra changes. This led to 2% lift in confirmed hires (key metric), and 14% lift in positive feedback ratio. Increased AI productivity (time it takes to build end-to-end model and productionize) by 7x (14 days→2 days).

  - Led a team of engineers for TF1→TF2 migration and PCv1→PCv2 migration of Job Recommendation system ranking, and contributed to design and implementation. This led to another increase in AI productivity by 2x (2 days→1 day).
  - Improved company diversity in Job Recommendations by feature engineering and joint optimization of relevance and company diversity. Defined new company diversity metrics for company diversity measurement.
- Owned, led, designed, and implemented Unified Offline Ranking EvaluatiOn (OREO) framework for LTSC (LinkedIn Talent Solutions and Careers) org, and integrated it with LinkedIn's ProML ecosystem. Onboarded users from external teams / orgs to OREO.
- Led and implemented multiple efforts on all aspects of Job Search Relevance, including query recommendations, candidate selection, and ranking.
  - Led Guided Search project for making query recommendations for Search starters and Inline query suggestions in Job Search product. Designed Unified Guided Search flow for Guided Search. Implemented and productionized Seq2Seq-LSTM model with Attention for generating query suggestions with online inference on free-form queries in Search starter and Inline Query Suggestion use cases. Inline Query suggestions with title→skill recommendations resulted in 1.5% lift in click rate@k.
  - Improved Job Search candidate selection by adding title-to-title expansion using title similarity via title Word2vec embeddings, high-quality skill-ID-based retrieval, and standardization-based skill-to-title expansion. These changes led to two successful MME models, which resulted in cumulative 1.4% lift in confirmed hires. Built a rule learning model to generate rule precedence, translated to Galene (based on Lucene).
  - Improved Job Search ranking with multiple changes, including: 1) Labeling strategy update (Premium / Basic job applies are equal-labeled), 2) Feature Engineering using Frame (Feature Management System): Added searcher-job skill-coverage percentage features to be used in ranking, 3) Using member / job skill / title topic embeddings in L2 ranking. Cumulatively, these changes resulted in 1.5% lift in confirmed hires.

Machine Learning Engineer 2016 - 2018
Team: Talent Relevance

- Led a team of engineers and researchers in Contextual Search project for Recruiter product.
  - Designed the offline and online architecture for Top-N skill recommendations for Contextual Search, led software engineers, and launched the product.
  - Co-designed the system for Candidate Recommendations with Query Generation and launched the product.
  - Improved Recruiter Search ranking using Contextual skill match features and topic vector match features via LDA. Designed and implemented offline and online pipeline for generating these features. Launched ranking models with contextual skill match features, and filed a patent for topic vector match features.
- Improved Recruiter Search ranking using various ranking models.
  - Designed, implemented, and productionized Generalized Linear Mixed (GLMix) models for Recruiter Search ranking with Learning-to-Rank features, XGBoost model scores and tree interaction features, with a 4-people team of engineers. This model led to 8.5% / 5% / 2% lift in InMail Accept@1/@5/@25 respectively, compared to baseline XGBoost model. Published and presented this work in WWW'19: Entity Personalized Talent Search Models with Tree Interaction Features.

  - Contributed to the design and implementation of an end-to-end deep learning model building pipeline for Recruiter Search using Tensorflow, Keras, Spark, Scala.
- Contributed to Recruiter Search ranking model training flow improvements.
  - Designed and implemented an end-to-end Offline Feature Engineering Workflow which allows users to quantify the impact of adding new features to Recruiter Search Modeling pipeline with hypothesis tests on the key evaluation metric.
  - Contributed to the migration of Recruiter Search label generation pipeline from Hadoop / Pig stack to Spark / Scala stack.
- Designed and implemented a Question Answering / Preference Elicitation system for Job Posting Flow, with adaptive question selection using entropy maximization after each question.
- Published 3 articles (SIGIR'2018, CIKM'2018, WWW'2019), filed 10 patents, and published 1 blog post on The AI Behind LinkedIn Recruiter Search and Recommendation Systems.

- **Research Scientist, Amazon** 2013 - 2016
  Research Scientist Seattle, WA
  Team: Consumer Marketing Analytics
  - Designed, built, and productionized ranking models for multiple business lines, product categories, channels, programs.
    - Product category purchase propensity models in 11 marketplaces using weighted logistic regression, which led to 20% lift in purchase rates and revenue overall.
    - Channel response propensity models (email / mobile) using weighted logistic regression, which are used in conjunction with other propensity models.
    - Mobile channel acquisition and engagement models, Amazon Mobile shopping app download propensity models for mobile acquisition. Live tests with mobile-adjusted product category purchase propensity models, compared to Mobile App First Sign-in (FSI) rate of 0.3% for universal control set, targeting Mobile shopping app download propensity model based segments with BAU offer and 5$ incentive offer, returned 413% and 1476% incremental lift on FSI rate, respectively.
    - Custom Audience Targeting: Targeting Amazon customers on Facebook, using product category purchase propensity models and deal seeker propensity models. In live tests, campaigns targeting product category segments drove 23.4% E%O, and campaigns targeting deal seeker segments along with product category segments drove 14.67% E%O, which are the largest ROI the social channels has seen at Amazon at the time.
    - Prime Free-Trial program sign-up propensity models using thresholded model ensembles. Using two different ranking methods in live experiments, customer segments selected based on Prime Free-Trial sign-up propensity models returned 94% and 74% higher Prime Free-Trial program sign-up rates compared to the baseline sign-up rate of 7%.
    - Various channel usage propensity models and customer life-cycle propensity models, including Kindle Cross-Platform reader download propensity models, Google channel reliance models, Amazon Attrition propensity models.
  - Designed, built, and productionized causal propensity models for product lines and channels.
    - Designed, built, and productionized causal inference models to calculate the differential propensity of a customer to take an action after targeting. Various applications include direct mail targeting for streaming Prime Instant Video,

making a purchase from Fashion product categories, signing up for Amazon Business program. Live tests for PIV streaming show that incremental models return a lift of 32% and 121% on the percentage of streamers respectively, compared to overall response models and random targeting.

  - Designed and built causal inference models with importance weighting, in order to measure the differential propensity of a customer to take an action after targeting in case of treatment set selection bias. In live experiments, bias-corrected uplift models returned higher incremental response rate compared to treatment model and biased uplift model. Filed a patent.

- Helped internal customers onboard to Predix, an automated predictive modeling platform, by helping them with model target generation, model building, and model evaluation.
- Organized and held Machine Learning Talk Series for Consumer Marketing organization as a monthly recurring event.
- Published articles, presented posters, gave talks in Amazon internal machine learning conferences.

- **Software Developer, Bank of America Merrill Lynch** 2012 - 2013
  Software Developer New York, NY
  Team: Equity Linked Technology
  - Designed and implemented software for processing TESS real time feeds within trading systems.
  - Designed and implemented a connector between Access Request Management (ARM) and RAM database.

- **Rensselaer Polytechnic Institute** 2006 - 2012
  Research Assistant Troy, NY
  - **TB-Insight**: http://tbinsight.cs.rpi.edu
    Supervisor: Prof. Bulent Yener (PI: Prof. Kristin Bennett)
    - UBF: Developed the Unified Biclustering Framework (UBF) to find host-pathogen associations among *M. tuberculosis* complex strains and TB patients.
    - Developed an algorithm to find the mutation history in the DR region of *M. tuberculosis* complex and found topological attributes of the resulting phylogenetic tree. Built two new models for mutation length frequency prediction: Starting Point Model (SPM) and Longest Block Model (LBM).
    - TCF: Developed the Tensor Clustering Framework (TCF) to cluster *M. tuberculosis* complex strains into coherent groups using multiple-biomarker tensors.
    - TB-Vis: Designed and implemented a visualization program for pathogen and host analysis of tuberculosis. The visualizations include: 1) Spoligoforests which display spoligotype evolution using various distance measures of genomic data of *M. tuberculosis* complex, 2) Host-pathogen maps which display patient characteristics classified by the genotype of *M. tuberculosis* complex which infects the patients. Spoligoforests are used in TB-Lineage tool at http://tbinsight.cs.rpi.edu/about_tb_lineage.html . Both spoligoforests and host-pathogen maps are used in the design of interactive version of TB-Vis at http://tbinsight.cs.rpi.edu/about_tb_vis.html .
  - **Enron Email Classification Into Topics**
    Supervisor: Prof. Sibel Adali, Prof. Boleslaw Szymanski
    - Classified email messages of Enron into topics using the Latent Dirichlet Allocation. Described new metrics for classification assessment of email distribution into topics.

- **MetPetDB**: http://metpetdb.rpi.edu
  Supervisor: Prof. Sibel Adali, Prof. Boleslaw Szymanski
  - Contributed to the design and implementation of the client-server architecture of a database for Metamorphic Petrology.

- **TUBITAK (Scientific and Technological Research Council)** Summer 2005
  Software Engineer Intern Ankara, Turkey
  - Designed and implemented a program which calculates the salaries of the employees using their entrance and exit time to/from the building.

- **TRT (Turkish Radio and Television)** Summer 2004
  Software Engineer Intern Ankara, Turkey
  - Contributed to the implementation of user interface of Eurovision Song Contest website.

Teaching

- **Teaching Assistant**
  Rensselaer Polytechnic Institute
  - Introduction to Algorithms Fall 2010
  - Computer Science II: Data Structures Spring 2008
  - Computer Organization Spring 2007
  - Artificial Intelligence Fall 2006
- **Undergraduate Teaching Assistant**
  Bilkent University
  - Discrete Mathematics Spring 2006

Patents

- R. Ramanath, G. Polatkan, Q. Guo, C. Ozcaglar, K. Kenthapadi, S. Geyik. **Unsupervised Learning of Entity Representations Using Graphs**. *Issued*, 2021.
- C. Ozcaglar, V. Dialani, S. Smoot, S. Geyik, A. Nair. **Feature Selection Impact Analysis for Statistical Models**. *Issued*, 2021.
- C. Ozcaglar, S. Geyik, P. Sharma, B. Schmitz, E. Buchanan. **Entity-Level Search Models with Tree Interaction Features**. *Issued*, 2021.
- C. Ozcaglar, R. Wu, J. Yang, A. Gupta, A. Nair. **Generalized Linear Mixed Models (GLMix) for Improving Search**. *Issued*, 2020.
- C. Ozcaglar, R. Ranjan, V. Parimi. **Uplift Modeling with Importance Weighting**. *Issued*, 2020.
- R. Ramanath, G. Polatkan, Q. Guo, C. Ozcaglar, K. Kenthapadi, S. Geyik. **Techniques for querying user profiles using neural networks**. *Issued*, 2020.
- R. Ramanath, G. Polatkan, Q. Guo, C. Ozcaglar, K. Kenthapadi, S. Geyik. **Generating candidates for search using scoring/retrieval architecture**. *Issued*, 2020.
- C. Ozcaglar, K. Kenthapadi. **Detecting Anomalous Candidate Recommendations**. *Filed*, 2019.
- K. Kenthapadi, C. Ozcaglar. **Feedback-based Update of Candidate Recommendations**. *Filed*, 2019.
- C. Ozcaglar, G. Borje, S. Geyik, G. Gulati, K. Thakkar. **Contextual Search Ranking Using Entity Topic Representations**. *Filed*, 2019.
- S. Wakankar, M. Meng, C. Ozcaglar, V. Abdrashitov. **Data Selection Based on Career Transition Embeddings**. *Filed*, 2019.
- P. Cheung, E. Buchanan, C. Liao, D. Boyd, G. Gulati, F. Or, C. Ozcaglar. **Utilizing Search Facets Based on Project Context**. *Filed*, 2018.

Conference/ Workshop Articles

- C. Ozcaglar, S. Geyik, B. Schmitz, P. Sharma, A. Shelkovnykov, Y. Ma, E. Buchanan. **Entity Personalized Talent Search Models with Tree Interaction Features**. *WWW*, 2019.
- R. Ramanath, H. Inan, G. Polatkan, B. Hu, Q. Guo, C. Ozcaglar, R. Wu, K. Kenthapadi, S. Geyik. **Towards Deep and Representation Learning for Talent Search at LinkedIn**. *CIKM*, 2018.
- S. Geyik, Q. Guo, B. Hu, C. Ozcaglar, R. Wu, K. Kenthapadi. **Talent Search and Recommendation Systems at LinkedIn: Practical Challenges and Lessons Learned**. *SIGIR*, 2018.
- S. Salem, C. Ozcaglar. **MFMS: Maximal frequent module set mining from multiple human gene expression datasets**. *KDD International Workshop on Data Mining in Bioinformatics (BIOKDD)*, 2013.
- C. Ozcaglar, A. Shabbeer, N. Kurepina, B. Yener, K. P. Bennett. **Data-driven insights into deletions of *Mycobacterium tuberculosis* complex chromosomal DR region using spoligoforests**. *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2011.
- K. P. Bennett, C. Ozcaglar, J. Ranganathan, S. Raghavan, J. Katz, D. Croft, B. Yener, A. Shabbeer. **Visualization of tuberculosis patient and *Mycobacterium tuberculosis* complex genotype data via host-pathogen maps**. *IEEE BIBM Workshop on Computational Advances in Molecular Epidemiology*, 2011.
- M. Aminian, A. Shabbeer, K. Hadley, C. Ozcaglar, S. Vandenberg, K. P. Bennett. **Knowledge-based Bayesian network for the classification of *Mycobacterium tuberculosis* complex sublineages** *ACM Conference on Bioinformatics, Computational Biology and Biomedicine (BCB)*, 2011.
- C. Ozcaglar, A. Shabbeer, S. Vandenberg, B. Yener, K. P. Bennett. **Examining the sublineage structure of *Mycobacterium tuberculosis* complex strains with multiple-biomarker tensors**. *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2010.
- A. Shabbeer, C. Ozcaglar, M. Gonzalez, K. P. Bennett. **Optimal Embedding of Heterogeneous Graph Data with Edge Crossing Constraints**. *NIPS Workshop on Challenges of Data Visualization*, Whistler, BC, 2010.
- J. M. Pyle, F. S. Spear, S. Adali, B. K. Szymanski, S. Pearce, A. Waters, Z. Linder, C. Ozcaglar. **MetPetDB: The Unique Aspects of Metamorphic Geochemical Data and Their Influence on Data Model, User Interface and Collaborations**. *Geological Society of America Abstracts with Programs*, Vol. 39, No. 6, 2007.

Journal Articles

- S. Salem, C. Ozcaglar. **Hybrid coexpression link similarity graph clustering for mining biological modules from multiple gene expression datasets**. *BioData Mining*, 2014.
- K. P. Bennett, C. Ozcaglar, J. Ranganathan, S. Raghavan, J. Katz, D. Croft, B. Yener, A. Shabbeer. **TB-vis: Visualizing TB patient-pathogen relationships**. *Tuberculosis*, 2013.
- C. Ozcaglar, A. Shabbeer, N. Kurepina, N. Rastogi, B. Yener, K. P. Bennett. **Inferred spoligoforest topology unravels spatially bimodal distribution of mutations in the DR region**. *IEEE Transactions on NanoBioscience*, 2012.
- C. Ozcaglar, A. Shabbeer, S. Vandenberg, B. Yener, K. P. Bennett. **Epidemiological models of *Mycobacterium tuberculosis* complex infections**. *Mathematical Biosciences*, 2012. (*Featured as the most downloaded article of Mathematical Biosciences journal in March-May 2012*).

- A. Shabbeer, L. Cowan, C. Ozcaglar, N. Rastogi, S. L. Vandenberg, B. Yener, K. P. Bennett. **TB-Lineage: an online tool for classification and analysis of strains of *Mycobacterium tuberculosis* complex**. *Infection, Genetics and Evolution*, 2012.
- A. Shabbeer, C. Ozcaglar, B. Yener, K. P. Bennett. **Web tools for molecular epidemiology of tuberculosis**. *Infection, Genetics and Evolution*, 2012. (*Featured as the most downloaded article of Infection, Genetics and Evolution journal as of December 2011*).
- C. Ozcaglar, A. Shabbeer, S. Vandenberg, B. Yener, K. P. Bennett. **Sublineage structure analysis of *Mycobacterium tuberculosis* complex strains with multiple-biomarker tensors**. *BMC Genomics*, 2011.

Technical Reports

- A. Shabbeer, C. Ozcaglar, K. P. Bennett. **Crossing minimization within graph embeddings**. *arXiv*, 2012.
- C. Ozcaglar, B. Yener, K. P. Bennett. **Host-pathogen association analysis of tuberculosis patients via Unified Biclustering Framework**. Rensselaer Polytechnic Institute. TR-12-05, 2012.
- C. Ozcaglar, A. Shabbeer, S. Vandenberg, B. Yener, K. P. Bennett. **A clustering framework for *Mycobacterium tuberculosis* complex strains using multiple-biomarker tensors**. Rensselaer Polytechnic Institute. TR-10-08, 2010.
- F. S. Spear, J. M. Pyle, S. Adali, B. K. Szymanski, A. Waters, Z. Linder, C. Ozcaglar, S. O. Pearce. **MetPetDB: A database for metamorphic geochemistry**. Rensselaer Polytechnic Institute. TR-08-14, 2008.

Theses

- **Algorithmic Data Fusion Methods for Tuberculosis**, *Ph.D. thesis*, Rensselaer Polytechnic Institute, 2012.
- **Classification of Email Messages into Topics Using Latent Dirichlet Allocation**, *M.S. thesis*, Rensselaer Polytechnic Institute, 2008.

Poster Presentations

- C. Ozcaglar, S. Geyik, B. Schmitz, P. Sharma, A. Shelkovnykov, Y. Ma, E. Buchanan. **Entity Personalized Talent Search Models with Tree Interaction Features**. *WWW*, 2019.
- C. Ozcaglar, B. Yener, K. P. Bennett. **UBF: Unified Biclustering Framework**. New York Academy of Sciences (NYAS) 7th Annual Machine Learning Symposium, NYC, 2012.
- C. Ozcaglar, A. Shabbeer, S. Vandenberg, B. Yener, K. P. Bennett. **A clustering framework for *Mycobacterium tuberculosis* complex strains using multiple-biomarker tensors**. *RPI-NSF Workshop on Multiscale Modeling of Complex Data*, Troy, NY, 2011.
- C. Ozcaglar, A. Shabbeer, S. Vandenberg, B. Yener, K. P. Bennett. **Insights into Camin-Sokal Parsimony and evolution of spoligotypes via spoligoforests**. *New York Academy of Science (NYAS) Symposium on Imaging, Visualization and Simulation: New Tools for Technology and Healthcare*, NYC, 2011.
- C. Ozcaglar, A. Shabbeer, S. Vandenberg, B. Yener, K. P. Bennett. **Multiple-biomarker tensor analysis for tuberculosis lineage identification**. *NIPS Workshop on Tensors, Kernels and Machine Learning*, Whistler, BC, Canada, 2010.
- C. Ozcaglar, B. Yener, A. Shabbeer, M. Aminian, K. P. Bennett. **A clustering framework for *Mycobacterium tuberculosis* complex strains using multiple-biomarker tensors**. *New York Academy of Science (NYAS) 5th Annual Machine Learning Symposium*, NYC, 2010.
- C. Ozcaglar, B. Yener, A. Shabbeer, M. Aminian, K. P. Bennett. **Examining sublineage structure of *Mycobacterium tuberculosis* complex strains with multiway modeling**. *Eigenvector University*, Seattle Washington, 2010. (Best poster award)

Talks

- **MFMS: Maximal frequent module set mining from multiple human gene expression datasets**. *ACM SIGKDD International Workshop on Data Mining in Bioinformatics (BIOKDD)*, Chicago, 2013.
- **TCF: Tensor clustering framework on multiple-biomarker tensors**. *Bogazici University, Middle East Technical University, Bilkent University*, 2012.
- **Data-driven insights into deletions of *Mycobacterium tuberculosis* complex chromosomal DR region using spoligoforests**. *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Atlanta, 2011.
- **Examining the sublineage structure of *Mycobacterium tuberculosis* complex strains with multiple-biomarker tensors**. *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Hong Kong, 2010.
- **Extracting Associations from Activities**. *NSF Cyber-Enabled Discovery and Innovation Symposium*, Troy, NY, 2007.

Awards

- Data All Star Team Award at LinkedIn, 2021.
- Transaction Risk Management Systems Modeling Hackathon at Amazon, first place, 2015.
- Student Travel Award to attend IEEE BIBM 2010.
- Full scholarship awarded by Rensselaer Polytechnic Institute for graduate study, 2006 - 2012.
- Full scholarship awarded by Bilkent University for undergraduate education, 2002 - 2006.
- Top 0.01% in nationwide University Entrance Exam among 1.5 million candidates, 2002.
- Ranked 1st in the Mediterranean Region in 8th and 9th Turkish Mathematics Olympiad, 2000, 2001.
- Bronze medal in 3rd and 4th Turkish Secondary School Mathematics Olympiad, 1998, 1999.

Activities

- **Program Committee Member:** CIKM 2019
- **Reviewer for Conferences and Journals:** KDD (2021, 2018), CIKM (2019), ACM Transactions on Algorithms (2013), Machine Learning (2017), IEEE Transactions on Knowledge and Data Engineering (2013, 2017), Computational Intelligence (2012, 2013), Network Modeling Analysis in Health Informatics and Bioinformatics (2011), Computational and Mathematical Methods in Medicine (2013).
- **Member:** IEEE, ACM.

Affiliations

- Treasurer of Turkish Student Association at RPI, 2011-2012.
- Member of Building Planning Committee, Computer Science, RPI, 2011-2012.
- RPI School of Science Graduate Council, Computer Science Representative, 2010-2011.
- Member of Graduate Recruiting Committee, Computer Science, RPI, 2007-2008.