DR. OSCAR RICARDO MOLL

Curriculum Vitae

Cambridge, MA — orm@csail.mit.edu — oscar-moll.com — linkedin.com/in/oscarmoll —

EDUCATION

Doctor of Philosophy in Computer Science (PhD) 2015-2023
MIT Department of Electrical Engineering and Computer Science
Thesis Title: Efficiently Searching for Objects Within Large Collections of Images and Video.
Minor in entrepreneurship.
Blockcert: https://credentials.mit.edu/certificate/048c651fc6ae5dce8c9e6242be2a5121

Master of Engineering in Computer Science (MEng) 2012
MIT Department of Electrical Engineering and Computer Science
Thesis Title: Database partitioning strategies for social network data

Bachelor of Science in Mathematics (BSc) 2011
Bachelor of Science in Computer Science (BSc) 2011
Massachusetts Institute of Technology (MIT)

EXPERIENCE

Contributor/Advisor
Pixeltable **Aug 2023 - Present**
Pixeltable is an open-source system that radically simplifies Machine Learning Engineering from research to deployment. It abstracts data management and computations behind a tabular interface, freeing MLEs to focus on their experiments without giving up control over the algorithms and custom code they run.

My contributions to the early development of Pixeltable focused on understanding user needs, reducing friction to get started using Pixeltable, designing and implementing parts of the Pixeltable interface, and mapping and demonstrating existing workflows, such as retrieval augmented generation, onto the Pixeltable Python interface. More specifically:

1. Implemented data import from Hugging-Face and Parquet datasets, and data export of Pixeltable DataFrames as PyTorch datasets to be used in model fine-tuning.
2. Integrated basic tooling for PDF document parsing and chunking, video image frame access.
3. Implemented Jupyter notebook integration for visualization of different Pixeltable data types such as documents and audio.
4. Streamlined the install process by packaging a complex PostgreSQL and pgvector dependency, enabling users to pip-install Pixeltable and try it in Google Colab.
5. Created Jupyter notebooks to demonstrate key target use cases such as RAG QA pipelines, which helped explain Pixeltable to potential design partners.
6. Prototyped API design alternatives to facilitate bringing existing code bases into Pixeltable
7. Interviewed design partners to understand their needs and workflows, helping guide Pixeltable development.
8. Supported design partners in a timely manner when they ran into problems or bugs.
9. Wrote tests, Github CI workflows, and reviewed code from other contributors.

Doctoral Research Assistant, Data Systems Group
MIT CSAIL, Cambridge, MA **Feb 2017 - Jun 2023**
I focused on the problems of managing and operating on databases of images and videos, my work develops new methods combining Machine Learning and Systems. Specifically:

1. Reducing the high compute costs of neural net-based image understanding AI algorithms, such as object detectors.
2. Reducing the high barrier to entry to applying AI techniques on your own data.
3. Reducing the high engineering cost of creating machine learning applications that work on your own data, for example by reducing the time it takes to find positive examples of rarer objects

To address the above challenges, I

1. Developed adaptive sampling techniques that exploit the uneven distribution of objects of interest and the high redundancy of data in video; by mathematically deriving estimates of rewards for different files or clips of video within large video datasets, and comparing their value against deep neural network based approaches. (ExSample paper in bibliography section)

2. Built systems to leverage multi-modal embedding models such as CLIP, as well as semi-supervised learning, and human-in-the-loop techniques to more quickly locate rare objects while keeping the burden on users lower than using state of the art Active Learning based approaches. (SeeSaw paper in bibliography section)

Beside the novel research contributions, day-to-day work as a research assistant involves a variety of engineering, implementation and communication skills, such as:

1. Implementing alternative approaches from the systems and machine-learning research literature, and prototyping my own new approaches. Implementations were mostly based on the existing Python data and ML ecosystems, eg. PyTorch, Pandas, Scikit-Learn, and Ray projects, as well as some web front-end work, to better visualize data and to test different human feedback labeling modalities.
2. Fine-tuning deep neural network models as part of many experiments, either as baselines or parts of solutions.
3. Designing and critically interpreting evaluation benchmark results, using standard benchmarks such as COCO or LVIS, or creating my own.
4. Diagnosing and debugging issues affecting the experimental results, from the selection of benchmark data itself, to data labeling errors, to implementation bugs. Error profiling of ML based systems.
5. Building scalable, parallel benchmarking pipelines that leverage large GPU clusters, in order to accelerate iteration and discovery.
6. Condensing interesting and useful results into plots and explanations, including dynamic visualization.
7. Building custom tools and experimenting with approaches make it easier to work with unstructured data.
8. Surveying emerging machine learning approaches and keeping up with the latest developments.

Postdoctoral Research Assistant, Data Systems Group
MIT CSAIL, Cambridge, MA **Jul 2023 - Mar 2024**
Built RAG pipelines to extract structured information, such as mathematical variables and their values, from descriptions in unstructured academic papers.
Guided a Masters level project on extracting data from plots via fine-tuning object detectors.

Machine Learning Engineer (PhD Intern) Tesla Autopilot, Palo Alto, CA **Jun 2017 - Aug 2017**
Developed a data-distribution change detection pipeline for Tesla Autopilot fleet telemetry (e.g., looking for sudden braking hot-spots in maps which negatively affect driver experience). Aggregated the full Tesla fleet data in S3 and applied statistical tests for data distribution changes. The pipeline was implemented as an Apache Spark Streaming Query via PySpark. Additionally, we implemented an interactive web dashboard to visualize pipeline outputs overlaid on a map that also allowed users to dig into the anomalies found.

Software Engineer (PhD intern)
Google, Mountain View, CA **Jun 2016 - Aug 2016**
Designed and implemented way to operate more efficiently on top of existing stored or transmitted protocol buffer data in binary form (a common binary data interchange format).
Implemented a small expression compiler using LLVM to directly evaluate basic programs over binary protocol buffer data, eliding unnecessary repeated parsing overhead and memory allocations, dispatch, and bounds checking overheads

Software Development Engineer II (Full-Time) **Sep 2014 - Jan 2015**
Software Development Engineer I (Full-Time) **Sep 2012 - Sep 2014**
Amazon Web Services, Seattle, WA
Designed and implemented data path components for the distributed storage layer that underpins the Amazon Aurora service. Some of this joint design work was awarded multiple US patents.
Designed and implemented an anti-entropy protocol: The mechanism keeps replicas in sync despite transient failures and brings blank-slate replicas up-to-date.
Mentored a summer intern, interviewed engineering candidates, reviewed code.

Software Engineer Intern
Twitter, San Francisco, CA **Jun 2011 - Dec 2011**
Designed and evaluated alternative DB partitioning methods for social graph data (part of MEng Thesis)
Developed distributed tracing of requests within Apache Cassandra nodes and integrated it with company monitoring tools.

PUBLICATIONS

Conferences: ACM SIGMOD (intl. conference on management of data), IEEE ICDE (intl. conference on data engineering), VLDB (very large databases)

Co-authors: Samuel Madden, Vijay Gadepally, Tim Kraska, Michael Cafarella, Favyen Bastani, Holger Pirk, Michael Stonebraker, Matei Zaharia

SELECTED CONFERENCE PAPERS

1. **Oscar Moll**, Manuel Favela, Samuel Madden, Vijay Gadepally, and Michael Cafarella. Seesaw: Interactive ad-hoc search over image databases. *Proc. ACM Manag. Data*, 1(4), dec 2023. `https://dl.acm.org/doi/pdf/10.1145/3626754`
2. **Oscar Moll**, Favyen Bastani, Sam Madden, Mike Stonebraker, Vijay Gadepally, and Tim Kraska. Exsample: Efficient searches on video repositories through adaptive sampling. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 2956–2968, 2022. `https://arxiv.org/abs/2005.09141`

CONFERENCE PAPERS

3. Holger Pirk, **Oscar Moll**, Matei Zaharia, and Sam Madden. Voodoo : A vector algebra for portable database performance on modern hardware. *PVLDB*, 9(14):1707–1718, 2016
4. Yehuda Afek, Alexander Matveev, **Oscar R. Moll**, and Nir Shavit. Amalgamated lock-elision. In Yoram Moses, editor, *Distributed Computing - 29th International Symposium, DISC 2015, Tokyo, Japan, October 7-9, 2015, Proceedings*, volume 9363 of *Lecture Notes in Computer Science*, pages 309–324. Springer, 2015
5. Favyen Bastani, **Oscar Moll**, and Sam Madden. Vaas: video analytics at scale. *Proceedings VLDB Endowment*, 13(12):2877–2880, August 2020

WORKSHOP/DEMO PAPERS

6. **Oscar Moll**, Aaron Zalewski, Sudeep Pillai, Samuel Madden, Michael Stonebraker, and Vijay Gadepally. Exploring big volume sensor data with vroom. *PVLDB (Demo paper)*, 10(12):1973–1976, 2017
7. **Oscar Moll**, Sam Madden, and Vijay Gadepally. Ad-hoc searches on image databases. In *Heterogeneous Data Management, Polystores, and Analytics for Healthcare*, pages 3–9. Springer Nature Switzerland, 2022
8. Holger Pirk, **Oscar Moll**, and Sam Madden. What makes a good physical plan?: Experiencing hardware-conscious query optimization with candomblé. In Fatma Özcan, Georgia Koutrika, and Sam Madden, editors, *Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016*, pages 2149–2152. ACM, 2016

PATENTS

9. Yan Valerie Leshinsky, James Mcclellan Corey, Samuel James McKelvie, Oscar Ricardo Moll Thomae, and Pradeep Jnana Madhavarapu. Efficient garbage collection for a log-structured data storage system, October 2019. US Patent and Trademark Office
10. Samuel James McKelvie, Benjamin Tobler, James Mcclellan Corey, Pradeep Jnana Madhavarapu, Oscar Ricardo Moll Thomae, Christopher Richard Newcombe, Yan Valerie Leshinsky, and Anurag Windlass Gupta. Individual write quorums for a log-structured distributed storage system, March 2019. US Patent and Trademark Office

AWARDS AND HONORS

Guatemaltecos Ilustres Orator - 2021
Seguros Universales, S.A.
Guatemala

Greylock X - Summer 2017
Greylock Partners
San Francisco, CA

Edwin Webster Fellowship - Spring 2015
MIT EECS
Cambridge, MA

TALKS

Efficiently Searching for Objects Within Large Collections of Images and Video
Cambridge, MA - May 2023
Thesis Defense. MIT Dept of EECS

SeeSaw: Interactive Ad-Hoc Image Searches Over Image Databases
Palo Alto, CA - May 2022
Stanford and University of Washington Workshop on Video Analytics

ExSample: Efficient Searches on Video Repositories through Adaptive Sampling
Kuala Lumpur, Malaysia (remote) - May 2022
IEEE International Conference in Data Engineering (ICDE)

The Case for Learned Sampling in Video Datasets
Cambridge, MA - January 2019
North East Database Day (NEDB)

Exploring Big Volume Sensor Data with Vroom
Munich, Germany - September 2017
International Conference on Very Large Data Bases (VLDB)

SERVICE TO PROFESSION

Reviewer
International Conference on Very Large Databases (VLDB) 2024

Reviewer
AI City Challenge 2021 CVPR Workshop

TEACHING EXPERIENCE

Database Systems (Fall 2015)
MIT

Mathematics for Computer Science (Spring 2011)
MIT

REFERENCES

Samuel Madden
MIT College of Computing Distinguished Professor of Computing
MIT CSAIL
madden@csail.mit.edu

Michael Cafarella
Principal Research Scientist
MIT CSAIL
michjc@csail.mit.edu

Marcel Kornacker
Co-CTO, Co-Founder
Pixeltable
marcel@pixeltable.com

Pierre Brunelle
CEO, Co-Founder
Pixeltable
pbrunelle@pixeltable.com