Richard Boeri Decal

USA · [redacted] · ✉ public@richarddecal.com

 crypdick · ≝ crypdick · in richarddecal · 🎓 GScholar · 🏠 richarddecal.com

Profile

I'm a self-starting Machine Learning Scientist/Engineer, proficient in full-lifecycle and full-stack projects, with 10 years of academic and industry experience. Having founded two ML departments at start-ups, I have a proven record of doing what it takes to ship useful products: iterating quickly, creating data curation systems, developing novel and SOTA solutions, architecting ML infrastructure, and scaling to the cloud. I have a proven record of quickly learning new domains (molecular biologist $\Rightarrow$ computational neuroscientist $\Rightarrow$ ML engineer) and am able to communicate complex technical concepts to all levels.

Style: 1) orient to customer needs, 2) design from first principles, 3) build in vertical slices.

Machine Learning Experience

Product Manager   San Francisco, CA
Anyscale   Jan. 2025 — Now

Anyscale is a cloud ML platform for scaling Ray Distributed workloads.

- Product manager for Ray Train and Ray Tune.

- Created end-to-end tutorials for scalable ML pipelines, spanning data ingestion, training, evaluation, and deployment.

Lead Machine Learning Engineer   Remote
Dendra Systems   Feb. 2020 — Jan. 2025

Dendra uses swarms of seeding drones to restore ecosystems and monitor biodiversity at scale. Founding ML lead. Full-stack, full life-cycle ML for scalable ecosystem restoration.

- Championed transformation from a services company to a ML product company using "Zone to Win" framework. My ML solutions have largely automated species identification tasks. This has transformed the business by increasing throughput, increasing profitability, reduced the 'cold start' problem with new biomes, enabled us to persue entire new markets.

- Initiative owner: training large computer vision models for species identification.

  - Bootstrapped end-to-end species ID stack: data processing, hyperparameter tournaments, training, evaluation, serving, monitoring.

  - Translated state-of-the-art self-supervised learning research into production to improve model robustness and reduce required labeled training data (`Pytorch`).

  - Researched, experimented, and productionized novel ML techniques: models, samplers, optimization functions, etc.

  - Developed few-shot learning models paired with an active learning data harvesting UI to automate new species and novel biomes in hours instead of months (`Gradio`).

- – Owner of ML roadmap (aligned with product roadmap & operations dept.). Established priorities, KPIs, and OKRs.
- – Hand-crafted data augmentations to make model robust to irrelevant features.

- Integrated species ID models into customer-facing platform and internal tooling.

  - – Conceived novel model-in-the-loop annotation tooling, accelerating insights delivery by over 80x.
  - – Devised model performance QC workflows to ensure we satisfy our SLAs.

- Set strategic vision for "data obsessed" ML and spearheaded our "data engine".

  - – Strategized overhaul of our data collection process to enable ML on long-tailed, open-world inference across thousands of target classes. Set business and system requirements, and system design for strategic labeling workflows (`PlantUML`).
  - – Implemented active learning methodologies to systematically harvest "high-leverage" data, preventing hallucinations on out-of-distribution data.
  - – Implemented novelty-maximizing data pruning to enable pareto-optimal (exponential) model scaling laws. Reduced training data by 60% while maintaining performance.
  - – Devised unsupervised "trip-wires" for detecting model failures in production. Integrated alerts into project tracker for strategic annotation team so that we can proactively fix the issue (`Jira`).
  - – Headed data curation tooling initiative (*C4 diagrams*). Point-person for external vendor assessment and selection.
  - – Created "ML University" lectures to educate ecologists on ML concepts and labeling best practices for high-quality data. Oversaw ML data collection team, developed rule-sets for data labeling and trained data annotation supervisor.
  - – Collaborated with ecologists to create model failure reports and gain intuition for model failures. Created data collection campaigns to patch biases in training data.
  - – Devised annotation QA and QC workflows: systematically identifying mislabeled and/or partially labeled samples to create "self-healing" training dataset.

- Scaling & Operational Excellence: Architected AWS-native cloud-scale infrastructure.

  - – Wrote distributed, scale-agnostic infrastructure for training, hyperparameter tuning, and inference (`Ray/Anyscale, AWS Batch`).
  - – Implemented Bayesian hyperparameter tournaments which aggressively kill underperforming trials, reducing training costs by 20x (`Ray Tune, HyperOpt, ASHA`).
  - – Identified bottlenecks and optimized throughput for multi-GPU jobs (`Grafana`).

- MLOps: Championed efforts to implement best practices for ML systems.

  - – Responsible for debugging model failures with paranoid programming, detailed chronicling, model interpretability algorithms (e.g. GradCAM, Transformer patch activiation maps), and heavy visualization of training dynamics.

Decal CV 2

- Responsible for full life-cycle of dataset and model artifacts, quality assurance: tracking artifact lineage, parameters for reproducibility (`MetaFlow`).
  - Devised sanity-checks to detect "silent failures" during model training.
  - Devised different stratifications for validating models, as well as validating specific data slices.
  - Enabled observability across pipelines (`Cloudwatch`, `Slackbots`, `UMAP`, `Sentry`). Reviewed metrics weekly to prevent customer-impacting incidents. Periodically reported the unit-economics of our labeling rates (`Jupyter`).
  - Operation Vacation: Led initiative to automate all workflows, including model training (`custom orchestrator`). Later, reimplemented as serverless to improve reliability and cost (`Step Functions, API Gateway`, λ, `EventBridge`).
  - Enforced code quality and correctness using pre-commit hooks, CI (`Bitbucket Pipelines`), ML sanity checks, property-based testing (`Hypothesis`), run-time validation (`Pandera`), design-by-contract (`beartype`).

Lead Data Scientist   Remote
PaceMate   Jan. 2019 — Dec. 2019

Pacemate monitors transmissions with bluetooth-enabled heart implants, identifying life-threatening arrhythmias and alerting emergency services. Founded ML division. Built end-to-end data processing and model training pipelines.

- Automated remote detection of cardiac arrhythmias in Internet-enabled heart implants using deep learning.

  - Developed processing pipelines for ECG data (`imbalanced-learn`, custom tools).
  - Working with cardiologists and software engineers to formulate business requirements (`YouTrack`).
  - Implemented state-of-the-art deep neural network for automated cardiac arrhythmia classification specifically tuned for the device implanted in a majority of our patients (`Keras`).
  - Created data labeling dashboard for electrophysiologists to review model predictions (`Plotly Dash`).

- Created dashboard to collate, explore, and summarize key insights from our electronic medical records.

  - Researched ML-assisted techniques for information extraction from extremely heterogeneous documents.
  - Wrote and scaled performant ETL pipelines (`SQL`, `PySpark`, `spaCy`).
  - Created dashboard to enable easy faceting and querying of EMR records to facilitate data-driven decision-making (`Plotly Dash`).
  - Created a report on our data inventory and trends in our data.

- Upheld SOC2 security standards with measures such as encryption at rest, traffic tunnelling, and instance hardening.

- Presented to the C-suite and met with potential investors.

Decal CV 3

Data Scientist   Sarasota, FL
New College of FL, F.A.R. Institute   Aug. 2018 — Dec. 2018

The Florence A. Rothman Institute supports innovation in medical data analysis. Semester-long master's capstone project supervised by Dr. Pat McDonald. Unpaid.

- Data-driven prediction of 30-day readmission using visit clustering.

    - `visit2vec`: reduce high-dimensional patient visit data into low-dimensional embeddings using a technique inspired by `word2vec` (`TensorFlow`).
    - Explored structure in patient visits data by clustering patient visits using t-SNE.

- Modeled patient trajectories on years of heart failure patients from Sarasota Memorial Hospital.

    - Clustered patients over time based on cardiac and non-cardiac chronic conditions (`SQL, Pandas, PySpark`).
    - Created network graphs characterizing interactions between multiple chronic conditions and heart failure and their effect on mortality (`NetworkX`)
    - Used finite state modeling to quantify interaction between chronic conditions and mortality (`PySpark, Numpy`).

Research Intern   Seattle, WA
Peng Lab, Allen Institute for Brain Science   June 2018 — Aug. 2018

The neuromorphology lab investigates the architecture of the brain at the population and single-cell level. Proposed a method that would automate the biggest bottleneck to high-throughput neural cell morphological analysis.

- Deep reinforcement learning for tracing neural structures in petabytes of noisy fluorescent microscope data.

    - Implemented proof-of-concept Deep Q Network using 3D convolutions to trace neural cell structures (`TensorFlow, rl-medical`).
    - Created simulation environment and reward system for training agents (`Matplotlib, OpenAI Gym`) based on manually traced microscope images.

Research Assistant   Seattle, WA
Fairhall Lab, University of Washington   Oct. 2014 — Jan. 2016

Computational neuroscience lab investigating the biophysics of neural cells. I developed agent-based dynamical models of mosquito thermal plume navigation behavior.

- Computed and visualized flight kinematic statistics and thermal sensing statistics using windtunnel flight data (`Numpy, Seaborn, scipy[interpolate, spatial, stats], sklearn, statsmodels`).

- Formulated biophysical models of mosquito thermonavigation: applied numerical optimization algorithms to fit model to experimental data (`scipy[optimize], Pandas`).

- Created animations of thermal plume navigation models (`Matplotlib 3D, MayaVi`).

Molecular Biologist Various
Various Pre 2014
Before transitioning to data science, I was formerly a molecular biologist.

Education

M.S. Data Science Sarasota, FL
New College of Florida Aug. 2017 — Dec. 2018

B.A., Chemistry/Biology (with honors) Sarasota, FL
New College of Florida Aug. 2007 — May 2011

Early admission (admitted 16 yrs old) Jupiter, FL
Harriet L. Wilkes Honors College Jul. 2006 — May 2007

Publications, Presentations, & Teaching

* 2024 Talk at *AI In Production Conference* on data curation.
* 2021 Invited talk *Ray Summit*: *How Ray and Anyscale Make it Easy to do Massive-scale ML on Aerial Imagery*. Accompanying blog post here.
* 2019 Seminar at *New College of FL*: *Remote Sensing of Cardiac Arrhythmia at Scale using Deep Learning*.
* 2019 Seminar at *Escuela Secundaria Tecnica de Torquinst*: *Intelligencia Artificial*.
* 2018 Classroom mentor for *Udacity's Intro to Programming Nanodegree: Python for Data Analysis Track* (1-on-1 tutoring, code reviews).
* 2015 UW Outreach: various educational events for students from low socioeconomic backgrounds.
* 2012 Two peer-reviewed journal articles in *Genetics* and *PNAS*, also presented as posters at three national conferences.
* 2007 Undergraduate honors thesis on RNA interference mechanisms in *C. elegans*.

Selected Awards & Grants

NCF Data Scholar 2017 — 2018
Full tuition waiver for master's program.

National Institutes of Health PA-12-149 Federal grant 2014 — 2016
Self-funded grant covering my salary and expenses at the UW Dept of Biophysics.

Florida "Bright Futures" Scholar 2007 — 2011
Merit-based scholarship. Full tuition.

Dubois-Felsmann Research Grant 2010 — 2011
Covered reagent costs for my thesis experiments & conferences.