

# Richard Boeri Decal

📍 USA · 📞 [redacted] · ✉️ [public@richarddecal.com](mailto:public@richarddecal.com)  
🐱 [crypdick](#) · 📄 [crypdick](#) · in [richarddecal](#) · 🏠 [richarddecal.com](#)

## Profile

Full-stack machine learning scientist with expertise in applying cloud computing and statistical modeling to real-world problems. Team lead experienced with 0-to-1 bootstrapping of ML products, data collection teams, and data engines (plus their maintenance and scaling). Former molecular biologist. First-generation Hispanic American.

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

## Machine Learning Experience

### Lead Machine Learning Engineer

[Dendra Systems](#)

Remote

Feb. 2020 — Now

*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.
- 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.
  - Created custom data augmentations to make model robust to irrelevant features.
  - Developed few-shot learning techniques to enable rapid bootstrapping insights to new species and novel biomes.
  - Owner of ML roadmap (aligned with product roadmap & operations dept.). Established priorities, KPIs, and OKRs.
- 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 the initiative.
  - 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), and heavy visualization of training dynamics.
    - 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,  $\lambda$ , 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

*PaceMate*

Remote  
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.

## Data Scientist

*New College of FL, F.A.R. Institute*

Sarasota, FL  
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

*Peng Lab, Allen Institute for Brain Science*

Seattle, WA

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](#).

### Classroom Mentor

*Udacity*

Remote

Dec. 2017 — May 2018

Provided 1-on-1 tutoring and code reviews for the *Intro to Programming Nanodegree: Python for Data Analysis Track*.

### Research Assistant

*Fairhall Lab, University of Washington*

Seattle, WA

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).

## Education

### M.S. Data Science

*New College of Florida*

Sarasota, FL

Aug. 2017 — Dec. 2018

### B.A., Chemistry/Biology (with honors)

*New College of Florida*

Sarasota, FL

Aug. 2007 — May 2011

### Early admission (admitted 16 yrs old)

*Harriet L. Wilkes Honors College*

Jupiter, FL

Jul. 2006 — May 2007

## Publications, Presentations, & Teaching

- Invited talk *Ray Summit '21: [How Ray and Anyscale Make it Easy to do Massive-scale ML on Aerial Imagery](#)*. Accompanying blog post [here](#).
- Invited talk *New College of FL '19: Remote Sensing of Cardiac Arrhythmia at Scale using Deep Learning*.
- Published [two peer-reviewed journal articles](#) which I presented as posters at three national conferences. Published an undergraduate honors thesis.
- Invited ML talk at *Escuela Secundaria Tecnica, Torquinst, Argentina '19*.
- Participated in several outreach programs for young students from low socioeconomic backgrounds.

## Selected Awards & Grants

<b>NCF Data Scholar</b> <i>Full tuition waiver for master's program.</i>	2017 — 2018
<b>National Institutes of Health PA-12-149 Federal grant</b> <i>Self-funded grant covering my salary and expenses at the UW Dept of Biophysics.</i>	2014 — 2016
<b>Florida "Bright Futures" Scholar</b> <i>Merit-based scholarship. Full tuition.</i>	2007 — 2011
<b>Dubois-Felsmann Research Grant</b> <i>Covered reagent costs for my thesis experiments &amp; conferences.</i>	2010 — 2011