Richard Boeri Decal

Asheville, NC, USA · [redacted] · public@richarddecal.com
crypdick · crypdick · richarddecal · AIjedi · 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.

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

Machine Learning Experience

Lead Machine Learning Engineer — Fully Remote
Dendra Systems — Feb. 2020 — Now

Founding ML lead. Full-stack, full life-cycle ML for scalable ecosystem restoration.

- Owner of ML roadmap (aligned with product roadmap & operations dept.). Established priorities, KPIs, and OKRs.
- Headed R&D and deployment of internally- and externally-facing ML products (`Pytorch`).
  - Refined product requirements with stakeholders for model-in-the-loop systems to scale species identification in UHR imagery.
  - Spearheaded KPI-driven development: KPI selection, tracking, review meetings, action plans, and retrospectives.
  - Researched, experimented, and productionized novel techniques: data augmentations, samplers, optimization functions, etc.
  - Translated state-of-the-art self-supervised learning (SSL) research into production-ready code.
  - Devised prediction QC workflows: systematic triggers for detecting model false positives and out-of-distribution data, preventing customer-impacting incidents.
  - Designed novel data collection tooling powered by reverse image search (`custom`).
- MLOps: Architected, developed, and monitor AWS-native cloud-scale ML workflows: data processing, hyperparameter tournaments, training, evaluation, serving, monitoring.
  - Responsible for full life-cycle of dataset and model artifacts, quality assurance: tracking artifact lineage, parameters for reproducibility (`MetaFlow`).
  - Optimized latency and throughput of elastically-scaling clusters (`S3, Ray/Anyscale`). Increased throughput of training and inference jobs >10x.
  - Enabled observability in our pipelines (`Cloudwatch, Slackbots, UMAP, Sentry`). Reported incidents weekly to prevent customer-impacting incidents.
  - Led initiative to automate all workflows (`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`).

- Set strategic vision for data-centric ML (aka "data obsessed").
  - Key strategist for company-wide transformation from a services company to a product-based ML company using the "Zone to Win" framework.
  - Created custom data augmentations to enable model to ignore irrelevant visual features.
  - Periodically reported the unit-economics of our labeling rates and strategized an overhaul of our data collection process (`Jupyter`).
  - Set business and system requirements, and system design for data collection process (`C4 diagrams`). Headed tooling overhaul. Point-person for external vendor assessment and selection.
  - Pioneered custom techniques and tools for:
    * data pruning to enable pareto-optimal (exponential) model scaling laws.
    * active learning methodology to systematically harvest "high-leverage" data.
    * few-shot learning to enable rapid bootstrapping in novel biomes.
  - Oversaw ML data collection team, developed rule-sets for data labeling and trained data annotation supervisor.
  - Devised annotation QA and QC workflows: systematically identifying mislabeled and/or partially labeled samples (`custom`).

Lead Data Scientist Fully Remote
PaceMate^{TM} Jan. 2019 — Dec. 2019

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 various senior stakeholders such as CEO, CTO, and CIO. Interviewed by potential investors.

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

Semester-long master's capstone project supervised by Dr. McDonald in partnership with the Florence A. Rothman Institute. 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

Summer research project wherein I 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`).
  - Generated and augmented training data from manually traced microscopy dataset.
  - Created simulation environment for and engineered reward signals for training agents (`Matplotlib, OpenAI Gym`).
  - Contributor to `rl-medical`, a `tensorpack` extension for anatomical landmark detection.

Classroom Mentor Fully Remote
Udacity Dec. 2017 — May 2018

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

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

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 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

- Invited talk *Ray Summit '21*: *How Ray and Anyscale Make it Easy to do Massive-scale ML on Aerial Imagery*.
- Invited talk *New College of FL '19*: *Remote Sensing of Cardiac Arrhythmia at Scale using Deep Learning*.
- Published three peer-reviewed journal articles. Published an undergraduate honors thesis. Presented at three national conferences.
- 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

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, fully covered college tuition.

Dubois-Felsmann Research Grant 2010 — 2011
Covered reagent costs for my thesis experiments, conference admission fees, & travel.