

RONALD DOKU

Senior Data Scientist & Machine Learning Engineer

San Francisco Bay Area · (202) 423-8549 · ronald.doku@gmail.com

[GitHub](#) | [Personal](#) | [LinkedIn](#) | [Portfolio](#) | [Research](#)

SUMMARY

Senior Data Scientist and Machine Learning Engineer with 4 years at Oracle Advertising building production NLP models, scalable ML pipelines, and monitoring for **600+ classifiers**. Applied A/B testing and causal inference to optimize advertising performance, contributing to **15% higher conversions and 12% lower customer acquisition costs**. Experience spans **advertising, healthcare, and cybersecurity**, with additional work in LLM/RAG systems, RLHF, and trustworthy AI. PhD in Computer Science focused on big-data security.

EXPERIENCE

Founding Engineer

Oct 2026 – Present

Afrusan Health · afrusan.com · Care-membership platform (Ghana pilot)

- One of four founding engineers building Afrusan's care-membership platform — a secure, auditable web platform connecting enrollment, consent, scheduled nursing work, physician review, referrals, and follow-up in one longitudinal care record.
- Setting the engineering baseline: server-side role/resource authorization, provenance-preserving clinical records, and idempotent nurse task offer/accept/assign workflows.

Postdoctoral Research Fellow, Clinical Informatics

Aug 2025 – Aug 2026

Howard University College of Medicine · Washington, DC

- Led reproducible clinical informatics research using NIH All of Us, MIMIC-IV, and NHANES, focusing on laboratory-data provenance, ML benchmark integrity, and decision-threshold reliability.
- Built data-provenance checks that flagged 1,482 of 19,470 (7.6%) LDL cholesterol results labeled as directly measured but exactly matching calculated values, highlighting potential labeling issues in clinical AI benchmarks.
- Designed a Data Transformation Service for Howard University Hospital to standardize clinical data from different systems while preserving its meaning, traceability, and reliability for downstream medical decisions.
- First-authored four peer-reviewed clinical studies (three published, one accepted in 2026); trained biomedical and public-health researchers in machine learning, deep learning, and AI model evaluation.

Founder & Lead Applied AI Researcher

Oct 2024 – Present

Haske Labs — Independent applied AI research lab · San Francisco Bay Area

- Built an LLM-free clinical search system using Bayesian inference and graph-based retrieval; improved top-5 precision by 63.8% relative to cosine retrieval in offline MIMIC-IV tests, with 2.65 ms mean retrieval latency.
- Prototyped confidence-gated clinical workflow routing: 38.3% of 10,000 retrospective MIMIC-IV encounters met simulated automation criteria, with remaining cases flagged for human review.
- Released governed-rank, an open-source Python library for constrained ranking; improved adverse-impact ratio from 0.77 to 0.92 in an offline COMPAS benchmark with accuracy guardrails. Provisional patent filed.
- Led GBP-Audit, an AI bias-correction framework using geometric analysis, statistical testing, and performance guardrails; reduced measured disparity by up to 73% without meaningful accuracy loss in successful offline corrections.

Machine Learning Engineer / Data Scientist

Oct 2024 – May 2025

BizzySync — Pre-Seed AI Venture (Product Validation) · Remote

- Developed an AI-powered partnership platform that matches small businesses and nonprofits and generates actionable collaboration plans; tested the prototype with approximately 40–50 prospective clients to validate the product ahead of seed fundraising.
- Built an LLM feedback and evaluation pipeline using PostgreSQL, Airflow, and human preference data to improve generated recommendations; offline experiments showed 15% improvement in engagement metrics and 20%+ higher plan approval rates.
- Reduced inference latency by ~10% through asynchronous processing with Celery; developed a TinyLLaMA fine-tuning workflow using reward modeling and reinforcement learning to improve response quality and reduce dependence on external LLM APIs.

Senior Data Scientist, ML/NLP & Experimentation

Aug 2020 – Jul 2024

Oracle Advertising · San Francisco, CA

- Applied Difference-in-Differences (DiD) to estimate the impact of an audience-targeting product launched without randomized testing; assessed pre-treatment trends and adjusted for advertiser characteristics.
- Used Propensity Score Matching (PSM) to reduce selection bias by matching advertisers on budgets, industry, and historical performance; evaluated treatment effects, covariate balance, and statistical significance.
- Designed and analyzed A/B tests of audience segments and targeting strategies, contributing to 15% higher conversions and 12% lower customer acquisition costs.
- Developed neural networks to predict which audiences would visit a website using Word2Vec features; explored Siamese networks with triplet loss to improve audience-targeting representations.
- Built SQL/PySpark pipelines with Airflow, reducing model training time by 25%; partnered with product, engineering, and account teams on client ROI dashboards and audience-targeting decisions.
- Implemented concept-drift monitoring across 600+ production classifiers; subsequent modeling improvements increased ad relevance by 7% and campaign efficiency by 10%.

Earlier engineering experience

IBM Research (Research Engineer Co-op, 2016–2017) · ConsenSys (Software Engineering Intern, 2018) · IBM Internship (Software Engineer Intern, 2014) · Dell EMC (Software Engineer Intern, 2013)

TECHNICAL SKILLS

ML / NLP / LLM: Python, SQL, PyTorch, scikit-learn, transformers, embeddings, RAG/Graph RAG, fine-tuning (LoRA, DPO), RLHF (PPO), ranking & retrieval

Evaluation & Experimentation: A/B testing, causal inference (PSM, DiD), calibration, uncertainty, selective prediction, concept-drift monitoring, offline benchmarking

Data & Infrastructure: PySpark, Airflow, AWS, Docker, MLflow, ETL pipelines, CI/CD

Healthcare Data: FHIR R4, EHR data pipelines, clinical data quality

SELECTED APPLIED AI PROJECT

- **Cookieless Personalization | Privacy-preserving ad/recommendation research** (haskelabs.com/projects/cookieless-personalization) — Session-level intent, confidence-gating, and policy-aware ranking without persistent user identifiers, evaluated across three public datasets.

SELECTED PUBLICATIONS & OPEN SOURCE

- Doku, R., et al. Detecting Friedewald-Substituted “Direct” LDL Cholesterol in a National Multi-Site EHR Cohort: Prevalence and Benchmarking Impact. *J. Applied Laboratory Medicine*, 2026.
- Doku, R., et al. Equation Disagreement as a Zero-Cost Uncertainty Signal for LDL-C Classification. *J. Applied Laboratory Medicine*, 2026.
- Doku, R., et al. LDL-C Equation Disagreement as a Threshold-Integrity Flag for Orthogonal Lipid Confirmation. *J. Clinical Lipidology*, 2026.
- Doku, R., et al. GBP-Audit: AI Bias Intervention with Built-In Safety-First Guardrails. *IEEE CogMI*, 2025. (Invited talk)
- Doku, R. The Confidence Gate Theorem: When Should Ranked Decision Systems Abstain?
- Doku, R., et al. Cybersecurity in Big Data Era: From Securing Big Data to Data-Driven Security. *IEEE Trans. Services Computing*, 2021.

Full list: [Google Scholar \(900+ Citations\)](#)

EDUCATION

Howard University — PhD, Computer Science (2020) · MS, Data Science (2016) · BS, Computer Science (2013)

Doctoral research: big-data security, decentralized data sharing, and scalable secure systems.