

Jarod Ferkin

jarodferkin.com | jtferkin@gmail.com | linkedin.com/in/jarod-ferkin-740189201 | github.com/herodthegreater |

Winter Garden, FL · Open to relocation

SUMMARY

Data Science senior at UCF (graduating Spring 2027) who builds models and ships them where they are used. Built **Pigeon**, a cockpit that **reverse-trains** a YOLOv8n detector (cut, measure, regrow) and a live site where anyone compresses a real image model in the browser. Puts LLMs into production with guardrails: **Stall Card**'s Claude intake pipeline and grounded question answering. Measures on the task, not a proxy: mAP@50, ROC-AUC, accuracy kept per byte.

EDUCATION

B.S. Data Science **University of Central Florida** Expected Spring 2027
Coursework: Algorithms for ML, Big Data Analytics, Statistical Methods 3, Matrix & Linear Algebra, Applied Time Series · A.A., Valencia College

MACHINE LEARNING PROJECTS

Pigeon · Model-Compression Research Instrument Live: herodthegreater.github.io/pigeon July 2026 – Present

- Built a PyTorch/FastAPI cockpit that **reverse-trains** a YOLOv8n detector toward one task (cut, measure, regrow): dependency-aware structured pruning (torch-pruning), fake quantization and fine-tune recovery, scored on real mAP@50. Three 7.2% cuts with regrowth kept **67% of native mAP@50 with 31% fewer parameters**, where one 20% cut kept 16%.
- Showed **precision has slack, structure does not**: a one-shot 5% channel cut drove bird mAP@50 to 0; with group-wise scales 8-bit cost 0.5% and 6-bit 9%; at 4 bits, FP4 levels kept 79% of native where integer levels with one scale per channel kept 16% at the same bit count.
- Found FLOP reductions did not lower Apple-GPU latency (a 30% prune cut MACs 35% while latency rose), so the cockpit reports compute and measured latency separately.
- Shipped a **live teaching site**: a MobileNetV3 bird classifier exported to ONNX with every weight as a graph input, so the page edits weights and re-runs it in **ONNX Runtime Web**. Per-tensor INT8 collapses it to 50%, per-channel holds 94.8%, and sensitivity-aware mixed precision reaches **5× smaller at 94.4%**; plus a guide to **43 compression methods** (31 run in the browser) and 10 coding lessons.

Python, PyTorch (MPS), Ultralytics YOLOv8, torch-pruning, FastAPI, ONNX, ONNX Runtime Web, NumPy, Pyodide, Astro

Stall Card (formerly barnManager) · LLM Features in Production Live: barnmanager.fly.dev June 2026 – Present

- Built an **AI intake pipeline** running Claude Haiku server-side in PocketBase (keys never reach the browser): a forwarded vet invoice, farrier confirmation, or show result (or a photo or PDF of one) is parsed, matched to the right horse and record type, and filed into a review queue for one-tap approval, with **confidence-gated auto-posting**.
- Built **"Ask the barn"**: plain-language questions answered by Claude Haiku from that barn's own records only, with anti-hallucination guardrails: it says so when the records don't answer, and a change it proposes is saved only after the user confirms.

Claude API (Haiku), TypeScript, React, PocketBase, SQLite, Docker, Fly.io

Predicting Song Failure on Spotify · Per-Genre Classification December 2025

- Framed a binary "failure" target (popularity < 30) on a **114K-track Spotify dataset** and built a separate model per parent genre with stratified splits; Random Forest reached a **5-fold cross-validated ROC-AUC of 0.63–0.88** by genre, well above a Naive Bayes baseline (0.58–0.62), and weak rock results traced to ambiguous labels, not the model.

Python, scikit-learn, pandas

AI-Native & Traditional Data Scraping Engine LLM vs. Rules March 2026

- Put the **Claude API** into a live pipeline that turns raw HTML into **structured JSON**, and built a 9-script pandas framework to **validate AI-native against traditional parsing** (Scrapy, BeautifulSoup, a custom rowspan matrix-expansion algorithm).

EXPERIENCE

DriveShack June 2024 – April 2026

Manager / Front-of-House Team Lead · Orlando, FL

- Grew revenue **20%+ in one fiscal year** as part of a leadership turnaround team, implementing **data-driven budgeting**.
- Bridged event-logistics teams, floor staff, and clients on six-figure events, and designed training programs that upskilled staff and removed reliance on external hiring.

SKILLS

ML & Models: PyTorch, torch-pruning, Ultralytics YOLOv8, ONNX / ONNX Runtime Web, scikit-learn, pandas, NumPy · quantization, structured pruning, fine-tune recovery, model evaluation (mAP@50, ROC-AUC, F1)
LLMs & Engineering: Claude API, structured extraction, grounded question answering · Python, TypeScript, SQL, R · FastAPI, React, PocketBase, SQLite, Docker, Fly.io