Pavel Machalek PhD

San Francisco, California, United States

pavel@machalek.io | linkedin.com/in/pavel-machalek-0152b8269 | (415) 890-6659 | github.com/ivanmladek

EXPERIENCE

SpaceKnow Inc.

AI Pretraining Infrastructure Lead — *2023–Present, San Francisco*

- Designed and maintained high-throughput, fault-tolerant data pipelines for LLM pretraining, integrating multi-modal data lakes and orchestrating distributed ETL workflows.
- Developed scalable scraping infrastructure using Playwright and RedisBloom for deduplication, supporting 1,000+ QPS web data acquisition with proxy rotation and robust error handling.
- Built automated preprocessing and feature extraction pipelines leveraging OCR (Nougat), NLTK, and language detection for large-scale scientific and technical document corpora.
- Ingested and processed both unstructured internet data and structured libraries of PDF books, enabling comprehensive and diverse training datasets for LLMs.

Spartacus Inc.

Chief Technology Officer — *2018–2023, San Francisco, CA*

- Built privacy risk scoring engine using advanced machine learning approaches with Python, analyzing 250+ data points per user across 100+ data brokers, achieving 92% accuracy.
- Designed and deployed intelligent data protection systems using TensorFlow and PyTorch for real-time threat detection and mitigation.
- Pioneered implementation of AutoGPT and BabyAGI frameworks (post-2023) for autonomous task execution in privacy protection workflows.

SpaceKnow Inc.

Co-Founder, Board Member, CEO — *2013–2018, San Francisco Bay Area*

- Led the development of SpaceKnow Satellite Activity Index using advanced machine learning, processing over 2 million square kilometers daily using Docker and TensorFlow.
- Architected geospatial analytics platform integrating computer vision systems for strategic influence on AI product direction.

The Climate Corporation

Senior Data Scientist — *2011–2013, San Francisco Bay Area*

- Reduced corn yield prediction error from ±15.2% to ±12.1% across 6.2M acres through advanced machine learning implementations.
- Built ETL pipelines processing 14TB/day of MODIS/Landsat data on 256-node distributed computing cluster.

NASA Ames Research Center

Senior Scientist — *2009–2011, Mountain View, CA*

- Improved Kepler photometry precision to 29ppm (from 42ppm) using innovative signal processing algorithms.
- Processed 1.7M star light curves using CUDA-accelerated pipelines with custom optimization techniques.

SUMMARY

Research engineer with deep experience in pretraining infrastructure, distributed data pipelines, and high-throughput scraping systems. PhD in Astrophysics from Johns Hopkins. Built and scaled ML infra for LLMs, scientific, and geospatial domains. Focused on robust, reproducible, and scalable data and ML systems.

CORE SKILLS

Data & ML Infrastructure
Distributed ETL, data lakes, cloud pipelines (AWS/GCP), Kubernetes, Spark, Airflow, Dask, batch/stream processing, S3, Parquet, BigQuery, MLflow,

Scraping & Web Data
Playwright, RedisBloom for deduplication, scalable proxy rotation, 1,000+ QPS scraping, robust error handling, BullMQ, high-volume web data acquisition

ML & Preprocessing
OCR (Nougat), NLTK, langdetect, feature extraction, tokenization, PyTorch, TensorFlow, ONNX, GGML, JAX, automated benchmarking, MLOps

EDUCATION

The Johns Hopkins University — **2004–2009**
PhD in Astrophysics
Baltimore, MD

PATENT **US10839211B2**
Multi-resolution multi-spectral deep learning based change detection for satellite images
Inventors: Michal Reinstein, Jakub Simanek, Pavel Machalek, Jan Zikes
Assignee: SpaceKnow Inc.
Filed: Aug 2017, Granted: Nov 2020