Daniel C. Ferreira

dcferreira.com | github.com/dcferreira

Vienna, Austria
daniel@dcferreira.com

Machine Learning Engineer

I'm a Machine Learning expert with **8 years of experience** in industry and academia, with a background in Mathematics. I'm mostly interested in **NLP and cybersecurity**, and consider myself a **generalist** who likes to do research, programming, devops, data engineering, and data science. I thrive in dynamic environments with motivated teams, and love learning new tools.

Technical Skills

Expert	:	Python, Linux, Git, Pandas, Transformers, TensorFlow, PyTorch, Keras, scikit-learn, NumPy, Docker, Databricks, Spark, Wireshark
Advanced	:	JavaScript, Scapy, GCP, AWS, Azure, SQL
Intermediate	:	MongoDB, R, Go, Rust, C, Photoshop, Inkscape

Experience

Freelance — *Vienna, Austria*
Machine Learning Engineer — Oct 2022 – Present

Cyan Security — *Vienna, Austria*
Machine Learning Engineer — Jun 2019 – Sep 2022

- Developed an ML system for website classification which reduced error rate by over 50% and worked in over 100 languages, using and ensemble of deep models (**BERT**) developed with **TensorFlow** and **transformers**
- Developed an ETL pipeline that fetched around 100k websites per day, extracted features, and classified them using **Spark** and **Databricks** on **AWS**
- Built a **webscraper** capable of fetching millions of pages per day via a **serverless architecture** on **GCP** and **Azure**
- Deployed multiple **NLP** and **image** ML models to production using **CI/CD pipelines** and **MLOps** best practices
- Containerized and deployed multiple ML inference models with **Docker**, **BentoML**, and **FastAPI**
- Deployed infrastructure for multiple external partners to label our data with **Label Studio** and **MTurk**
- Mentored a student developing a tool for detecting **DNS tunneling** activity
- Defined a unified **REST API** for delivering input/output to/from the in-house ML models
- Developed a tool to extract **Zeek** network features from **PCAP files** with Python

(Interim) Machine Learning Team Lead — Aug 2022 – Sep 2022

- Managed a team of 4 ML engineers during the team lead's parental leave
- Represented the company in the initial stage of a research project with 3 academic partners

Technical University of Vienna — *Vienna, Austria*
Researcher in the Big-DAMA project — Aug 2016 – Mar 2019

- Published 6 research papers in ML methods for **cybersecurity** and similar topics
- Developed a tool to visualize **network traffic flows** in 2D and aggregate them based on labels, using **Autoencoders**
- Launched and managed a public initiative for cataloging and categorizing **network traffic** related research papers
- Developed **Python library** for a random data generator for research on clustering algorithms
- Developed an ML tool to **generate images** of façades in different cities' styles, using **GANs**
- Collaborated with a team from Carnegie Mellon University in developing and publishing a framework for neural network architecture search (**AutoML**)
- Co-advised a student on a Machine Learning thesis

Priberam Labs — *Lisbon, Portugal*
Junior Researcher in the SUMMA project — Feb 2016 – Jul 2016

- Researched, generated, and published one of the first pre-trained **multilingual word embeddings**
- Developed a machine learning model for **Named-entity Recognition** in multilingual news articles and media
- Defined a general **REST API** for an H2020 project with 10 international partner organizations

Education

Instituto Superior Técnico — *Lisbon, Portugal*
MSc in Applied Mathematics, major in computation — Sep 2013 – Dec 2015
Thesis in Cross-lingual Text Classification (grade 19/20)

BSc in Applied Mathematics — Sep 2010 – Jul 2013

Selected Projects

Toxic News (*Python, MongoDB, GCP, HTML, JavaScript, Git*) GitHub

- Designed and deployed an end-to-end system using serverless architecture that fetches websites on a schedule, runs them through off-the-shelf ML models, and displays results in a static web page
- Developed a CI/CD pipeline for testing commits, pushing new versions to production, and updating the live website
- Designed a modern responsive website using HTML, JavaScript, and Tailwind, and deployed it on GitHub Pages

Tweet Fake (*Python, Flask, Docker, CoHere, Git*) GitHub

- Engineered prompts for an app that takes your tweets and generates new ones in your style
- Implemented Twitter's OAuth workflow in Flask
- Containerized the app with Docker and deployed it in my personal server

Infinity for Youtube (*TypeScript*) GitHub

- Developed and published a Chrome extension for better usability on YouTube

Deep Architect (*Python, Git*) GitHub

- Conceptualized and implemented an early version of a neural architecture search framework
- Collaborated in a research publication accepted at NeurIPS

City-GAN (*Python, PyTorch, Git*) GitHub

- Collected and curated data using Google Maps APIs
- Implemented a state-of-the-art conditional GAN to generate images of façades in PyTorch
- Wrote and published a research paper on arxiv

Traffic Flow Mapper (*Python, TensorFlow, Keras, Git*) GitHub

- Researched the use of autoencoders to visualize network traffic flows
- Implemented the autoencoders using TensorFlow and Keras
- Wrote a research paper accepted at IJCNN

MDCGenPy (*Python, Docker, Git*) GitHub

- Developed a data generation Python library that researchers can use to evaluate clustering methods
- Wrote a research paper accepted at Springer Journal of Classification

NTARC Database (*Python, JavaScript, Docker, JSON Schema, Git*) GitHub

- Lead a community effort to catalog research papers and used techniques in network traffic analysis
- Developed a suite of tools to facilitate the curation process, including an Electron app to add entries to the database, a Python tool to verify syntax and correctness of user inputs, and a Python library to interact with the database
- Wrote and published 2 research papers with findings from the database effort

Multilingual Joint Embeddings (*Python, Theano, Git*) GitHub

- Developed a ML model that reduced the state-of-the-art error rate by 40% in a cross-lingual news classification task
- Obtained state-of-the-art results in multiple cross-lingual classification tasks
- Published multilingual word embeddings for 12 different languages
- Wrote a research paper accepted at ACL

More projects at dcferreira.com/project/

Selected Publications

- Extreme Dimensionality Reduction for Network Attack Visualization with Autoencoders, IJCNN 2019
- Towards modular and programmable architecture search, NeurIPS 2019
- Jointly Learning to Embed and Predict with Multiple Languages, ACL 2016

More publications on Google Scholar