Gerard Casas Saez contact@gerard.space

Website: gerard.space GitHub: @casassg

## Summary

- Interested in building Machine Learning and Data products and infrastructure at scale.
- Work experience in Machine Learning infrastructure particularly on model training and hosting solutions.
- Contributed to several Open Source projects like Apache Airflow, Kubeflow, TFX and Triton Server.

## Experience

#### • Block Inc

L6 Machine Learning Engineer, September 2024 - Present

- Co-authored and launched initial version for automatic model monitoring, tracking input and output drift.
- Led support for hosting FAISS approximate nearest neighbors within our model hosting solution.
- Collaborated across teams to evaluate next generation model training platforms to scale up further.
- Mentored new team members for growth and continued development.

#### • Block Inc

L5 Machine Learning Engineer, March 2023 - September 2024

- Developed automatic inference request storage system for statistical analysis and model debugging.
- Co-authored initial version for multi-model hosting endpoints reducing cost by 50%.
- Co-author of Python SDK for automatic model deployment reducing deploy times to hours instead of days.
- Developed initial support for hosting LLM within Block.
- Presented in industry conference about scaling up model hosting at Block.
- Co-authored initial next-gen training stack evaluation with the creation of Slurm cluster in GCP as POC.

#### • Twitter/X

Senior Machine Learning Engineer, April 2022 - February 2023

- Co-authored and launched initial version of X Inc large scaling training cluster with Slurm and Ansible.
- Technically led team of 8 engineers on ML Pipelines team, defining roadmap and execution cadence.
- Led internal adoption of Kubeflow Pipelines for model retraining.
- Designed and implemented next-gen re-training pipeline components API for both TFX and KFP.
- Worked with industry collaborators from Spoify and Google for alignment on reusable components for automatic model re-training pipelines.

### • Twitter

Machine Learning Engineer II, October 2020 - April 2022

- Led adoption and migration to production model retraining scheduling framework on top of Kubeflow clusters.
- Led development for comprehensive automatic CI/CD systems.
- Developed automatic Python wheel extractor program to help develop code in interactive environments outside
  of the monorepo.
- Authored VSCode Twitter monorepo intergration for Python developers, speeding development time.
- Co-authored evaluation for next-gen ML Training Service on GCP.
- Taught internal class on improving model quality by setting up automatic model re-training.
- Collaborated with Ads team to launch Ads Click model in GCP with improved developer iteration, leading to faster model improvements.

### • Twitter

Machine Learning Engineer, September 2019 - October 2020

- Co-authored internal hyperparameter optimization framework for using Bayesian optimization with Ax.
- Supported ML research work for "Tuning Word2vec for Large Scale Recommendation Systems" (RecSys2020).
- Worked on evaluation of industry ML model automatic retraining solutions and eventual adoption of TFX for defining end-to-end ML pipelines.

### • Project EPIC, University of Colorado Boulder

Research Assistant, February 2017 - August 2019

- Analyzed 2000 tweets/minute of real time data by building Big Data pipeline using Spark, Kafka and Cassandra.
- Deployed system prototype with Kubernetes using custom microservices written in Go and Python.

#### • Twitter

Software Engineering Intern, May 2018 - August 2018

- Developed CLI to deploy Apache Airflow instances internally.
- Helped reduce ML model training and deploy time by 1/4 developing ML Workflows.

## • inLab FIB

Web Development Intern, March 2015 - January 2017

- Built REST API for faculty data Average usage of 25.000 daily accesses. (Django, Python)
- Reduced Oracle SQL query response time by 1/3 using non-blocking updates and materialized views.

• con terra GmbH Intern, Summer 2015

- Developed bike share routing app using OpenStreetMap open data and Citibikes API. (Javascript)

## Education

# University of Colorado Boulder

Boulder, CO

M.S., Computer Science (GPA 3.97)

August 2017 - August 2019

Courses: Analysis and Design of Algorithms, User Centered Design, Database Systems, Big Data Architecture,
 Machine Learning, Computational Genomics, Natural Language Processing.

#### Universitat Politècnica de Catalunya

Barcelona, Spain

B.S., Computer Science (8.8/10)

September 2013 - July 2017

Courses: Data Mining, Data Compression, Programming Languages, Software architecture, Databases,
 Parallelism, Project Management, Web architecture, Android, NoSQL databases, Reflective programming.

## Volunteering

• TFX Addons SIG

Core Contributor, May 2021 - March 2023

- Release manager from version 0.1.0 to date.
- Contributor and maintainer for project CI/CD system, XGBoostEvaluator, FeastExampleGen and others.
- Collaborator for ModelCardToolking migration mentoring Twitter team members to contribute to project.
- AIP-31: Airflow functional DAG definition / "Taskflow API" Core Contributor, January 2020 July 2020
  - Proposed and implemented improvement for Airflow DAG user experience adding explicit XCom data.
  - Released in Airflow 2.0.0 and used in one of the main tutorials for the project.
- HackCU, University of Colorado Boulder

Tech director, February 2017 - April 2019

- Co-led Colorado's biggest student hackathon with 600 attendees and a 15 people core team.
- HackUPC, Universitat Politècnica de Catalunya

Co-Director, November 2015 - September 2018

- Co-led Spain's biggest student hackathon with 500 attendees and a 20 people core team.
- Led web development team with 5 people for 2 years, developing the event website and registration platform.

#### Awards

- Google OSS Peer Bonus for contributions to TFX and TFX Addons (2022)
- Denver Government award winner at ETHDenver. (2018)
- Grant winner at MD5 Physical Cyber Hackathon. (2018)
- Balsells Fellowship. (2017-2019, two years of funding for graduate studies)
- Balsells Mobility program. (2017, five months of funding for research)
- 52North Innovation Incentive Award. (2013)

### **Publications**

- Jennings A., Gerard C., Kenneth M. A., Leysia P., and Rebecca M. Incorporating Context and Location Into Social Media Analysis: A Scalable, Cloud-Based Approach for More Powerful Data Science. *HICSS 52* (accepted).
- Gerard C. Social Media Analysis for Crisis Informatics in the Cloud MS Thesis

## Skills

- **Programming:** Java, Python, GoLang **Experience with:** Kubeflow Pipelines, Django, Apache Airflow, Apache Beam, Spark, Kafka, Kubernetes, gRPC, Triton Server, Slurm, GCP, AWS
- Native: Catalan, Spanish. Advanced: English.
- Hobbies: Skiing, Running, Photography, Roller hockey, Piano, Swim.