N/A
94123 San Francisco, CA (U.S.)
N/A
N/A
www.linkedin.com/in/fperezsorrosal
francisco-perez-sorrosal.github.io

Francisco Perez-Sorrosal

Princ. Research Engineer (Yahoo Inc.)

"Strive not to be a success, but rather to be of value."
— Albert Einstein

Profile and Objectives

I'm a trustful, goal-oriented, detail-focused professional with experience in both research and software engineering (also in both academic and non-academic environments), particularly in the fields of machine learning, artificial intelligence, and distributed systems. As a Research Engineer, I love exploring and implementing with cutting-edge technologies, particularly in AI/ML, to tackle complex challenges and foster innovation. As a Software Engineer, I can design, develop, and ensure the maintainability of robust and scalable big data applications, whether on-prem or cloud, always prioritizing high availability, scalability, performance, and reliability. I am adaptable, capable of taking initiative and leading efforts with a strategic approach when needed, or collaborating as a communicative, supportive, and empathetic team player to achieve shared goals. I thrive in open, dynamic, and multidisciplinary environments, integrating visionary and practical insights to deliver impactful results while empowering others and fostering collaboration.

Professional Experience

2023–Now **Principal Research Engineer**, *Knowledge Graph Science @ Yahoo Inc.*, Sunnyvale

YKG, *Dev/Test Dataset Imnprovement*, 2024, **Roles**: ML/AI Expert
Goal As dataset labels were semi-supervised generated, develop an LLM-based solution to re-assure the labeling accuracy of the dev and test datasets
Achievement Generative Entity Matching solution with Llama 2 to identify and fix misclassified examples in the dev/test sets

YKG, *Entity Reconciliation Model Training*, 2023-2024, **Roles**: Machine Learning/AI Expert
Developed and trained reconciliation models to dedup People & Creative Work meta-entities
Goal Improve current production metrics on both categories
Achievements 1) Overall increases of 22.23% and 16.11% in precision and recall for Person 2) Overall increases of 5% and 4% in precision and recall for Creative Works

YKG, *Entity Reconciliation Pipeline*, 2023, **Roles**: Machine Learning/AI Expert, MLOps
Developed an automated pipeline for training entity reconciliation models for YK, from data curation to inference testing
Goal/Achievement Devise a solution to reduce the current number of models maintained in production from $\tilde{3}0$ heterogeneous (heuristic-based, SVMs, Tree-based...) to just 4, which encompass the main 4 meta-entity types in YK

YKG, *Entity Reconciliation Inference*, 2023, **Roles**: MLOps
Architect a develop a PoC solution for inference in the cloud
Goal Explore a Ray-Serve based approach for putting into production the new YK models in the cloud

Jun–Dec 2022 **Principal Research Engineer**, *Yahoo Mail*, Sunnyvale

Science @ Mail, *Kamino Project - mail classification system*, 2022, **Roles**: Machine Learning/AI Expert

Goal Update the old production models with a new breed of deep learning based small distilled models suitable for running at scale. The models were classifying the incoming mail based on a new multilabel taxonomy created specifically for new use-cases required by the mail team.
Approach I built a deep learning based pipeline with Hugging Face to train/evaluate deep learning models performing knowledge distillation following the Teacher/Student approach. Starting first from limited human-labeled training data, we built a large/complex, more accurate model; then we used a large amount of teacher-labeled data to train lightweight student models for suitable for deployment. Produce two kinds of models: 1) online student models; faster but less accurate suitable for real-time classification tasks. 2) offline student model; slower but include richer information that provides better accuracy results.

Achievements

The resulting student models for an expanded taxonomy, more than doubling the number of deployable categories in the taxonomy while improving performance on existing categories in production by 3.1-5.9%

Apr–Sep 2022 **Principal Research Engineer**, *Digital Transformation Office @ Yahoo*, Sunnyvale

AI/ML Working Group, *AI/ML future strategy definition in the cloud*, **Roles**: ML/AI Expert, Strategy

Yahoo infrastructure was going to transition from its own datacenters and resources to work fully in the cloud. Scientist and Research Engineers training, running, and serving ML/DL models needed a modern set of tools to share datasets, models and reproducible experiments with colleagues, rapidly experiment with, and iterate on new models, and finally, publish and deploy them easily into pre-production/production environments.

Goal The Model Development group in particular had as mission to determine the requirements, standards, tools, and frameworks to enable ML/DL applications at scale across the company.

Achievements

- I led the subgroup on Model Development, although I participated actively in the remaining three (Data Management, Model Management, and Model Serving)
- A set of recommendations (prioritized and categorized in topics) and tasks to do (internal processes, external relations, tools, etc.) to guarantee that ML/DL practitioners could work with less friction in the new cloud environments.

2021–2022 **Principal Research Engineer**, *Ads @ Yahoo*, Sunnyvale

AI/ML Working Group, *AI/ML future strategy definition in the cloud*, **Roles**: ML/AI Expert, Strategy

In the advertising platform, as we were moving toward a cookie-less world, the ability to track users' online activities for behavior targeting was be drastically reduced, making contextual targeting an appealing alternative for advertising platforms. Using our experience in hierarchical multilabel classification in other contexts, we helped the Ads team to use it in category-based contextual targeting at Yahoo. We proposed and implemented a multilingual model that can accurately classify web pages into a hierarchical taxonomy (specifically, the Yahoo Interest Categories taxonomy) without crawling their content. **Goal** Transfer the knowledge and experience in taxonomy-based multi-label classification into Ads team and build a platform for training/evaluating the models

Achievements

- I helped the Ads team to develop a pipeline for training models for the task at hand
- *Multilingual taxonomic web page classification for contextual targeting at Yahoo* paper. In the 28th ACM SIGKDD Conference, 2022 **7 Citations**

2018–2023 **Principal Research Engineer**, *Science @ Content Platform @ Yahoo Inc.*, Sunnyvale

Science @ CAP, *Deep Learning-based Muiti-label Classification*, 2021, **Roles**: ML/DL Engineer, Research Engineer
Goal Modernize the production models and infrastructure using new deep learning models.
Achievements - A modern pipeline to train/evaluate multi-label, multi-class and binary classification problems
- A tool for plugin taxonomies for the multi-label configurations
- Train models with significant better metrics over the ones in production (+10% over baseline)
- Develop a serving pipeline based on NVDIA's Triton Server to deploy in production

Science @ CAP, *Scalable Few-shot Classification with Parallel Prefix Conditioning*, 2021, **Roles**: ML/DL Engineer, Research Engineer
Goal Modify a transformer architecture (e.g. BERT) to achieve fast and effective few-shot classification by prefixing multiple category labels to the input.
Method Class representations are encoded in parallel but produce independent binary labels for each input through a shared classification output layer.
Achievements - Experiments on the DBPedia dataset demonstrated improved few-shot performance over standard multi-class classifiers and a speedup over binarized formulations.
- Further analysis showed that the approach can scale to a large number of categories and may hold promise for zero-shot learning of unseen categories.

Science @ CAP, *Clickbait Classifier based on Transformers*, 2019, **Roles**: ML/DL Engineer, Research Engineer
Goal Improve the current SVM-based model in production using a BERT-based model
Achievements
- Improve the current classifier in production by 5%
- Devise an integration strategy for serving the model in production using the existing pipeline
- Extra: PoC of a generative-based clickbait classification approach using T5

Science @ CAP, *Hierarchical Transfer Learning for Multi-label Text Classification*, 2018, **Roles**: ML/DL Engineer, Research Engineer
Goal Propose a novel transfer learning based strategy where binary classifiers at lower levels in a hierarchy of classes are initialized using parameters of the parent classifier and subsequently fine-tuned on the child categories for the classification task
Achievement Paper published in ACL 2019
- *Hierarchical transfer learning for multi-label text classification.* In ACL, 2019 **127 Citations**

2015–2018 **Senior Research Engineer**, *Content Ingestion Platform (CAP) @ Yahoo Inc.*, Sunnyvale

CAP Team, *DL-based multi-class Classification for Content Ingestion*, 2018, **Roles**: ML/DL Engineer
- Experiment with LSTM/GRU-based models for multi-class classification using TF/Keras
- Explore transfer learning for multi-label text classification

CAP, *Machine Learning Content Classification Pipeline*, 2017, **Roles**: ML Engineer
- Build a pipeline for easy data ingestion, model training and evaluation based on SVMs for the Sieve platform
- k8s-make, a tool to harness the compute-power in Yahooś on-prem clusters to deploy a tailor-made Kubernetes cluster to parallelize the training of our SVM-based pipeline in a simple way

2015–2017 **Senior Research Engineer**, *Sieve @ Yahoo Inc.*, Sunnyvale

Sieve, *Twitter Firehose: At scale, ingestion streaming system for Tweets*, 2017, **Roles**: Distributed Systems Expert, Research Engineer
- Reimplement the new Twitter API
- Re-architect previous solution to be more performant yet keeping the backwards compatibility
- Collaborate with the Sports team for productization

Sieve, *Scalable Content Ingestion Platform*, 2016, **Roles**: Distributed Systems Expert, Research Engineer
- Move Omid as full open-source project into the ASF (https://github.com/apache/incubator-omid)
- Continue Supporting the Omid transaction manager project in production
- *Omid* presented at Hadoop Summit. 2016, San Jose, CA (USA) - *Omid, reloaded: scalable and Highly-Available transaction processing* paper. In USENIX FAST, 2017 **20 Citations**

Sieve, *Scalable Content Ingestion Platform*, 2015, **Roles**: Distributed Systems Expert, Research Engineer
- Work at multi-tenant content ingestion platform at Search organization
- Add High Availability to the Omid Transaction Manager for HBase
- Scale Omid in multi-core architectures
- Support Omid transaction manager project in production infrastructure for content ingestion platform
- Explore moving Omid as open-source project in the Apache Software Foundation (ASF)

2012–2015 **Research Engineer**, *Scalable Computing Group @ Yahoo Labs.*, Spain

Omid, *Transaction Manager for Big Datastores*, 2014, **Roles**: Distributed Systems Expert, Research Software Engineer
- Re-architect original codebase/Implementation of new features
- Support integration in Sieve project
- Presentation and poster at Yahoo's Techpulse Conf. 2014

Edentity/Pachiderm, *Ubiquitous content access and management of personal data (images, video...) held on 3rd party services (Flickr, GDrive, Dropbox...).*, 2014, **Roles**: Distributed Systems Expert, Research Engineer
- Design and implementation of a scalable synchronization module
- REST API definition and implementation for each module defined (User Registration, Search, Content Fetching)

RiddlR, *Percolator-like incremental processing system.*, 2013, **Roles**: Distributed Systems Expert, Research Engineer
- Prototype and example application implementation
- Poster at Yahoo's Techpulse Conf. 2013

CumuloNimbo, *7th European Framework Programme (FP7-257993)*, 2012–2013, **Roles**: Distributed Systems Expert, Research Engineer
- Design and implementation of a prototype of an incremental processing system for Big Datastores
- Provide durability to HBase through BookKeeper
- Represent Yahoo in project meetings and evaluation

2011–2012 **Software Architect**, *Lumata*, Spain
SONY Socialife application. High Scalable Big-Data Backend Platform for web and social content ingestion and aggregation

Giddra Project, 2012, **Roles**: Distributed Systems Expert, Architect
- Architectural definition of the platform
- Team coordination
- REST API definition for allowing clients to access the backend

2010–2011 **Freelance Software Architect/Engineer**, *Local Greenhouse Cooperative in Zaragoza*, Spain
Analysis, design and Ruby/Rails implementation of a web application to manage the different domains of a greenhouse farm. LoC $\simeq$20.000

2003–2010 **Software Architect/Research Engineer**, *School of CS at Univ. Politécnica de Madrid (UPM)*, Spain
Participation in European and national research projects that included the analysis, design, implementation and testing of middleware architectures and applications. *Detailed achievements in projects*:

NEXOF-RA: NESSI Open Framework - Reference Architecture, *7th European Framework Programme (FP7-216446)*, 2008–2010, **Roles**: Software Architect
- Contribution to a reference architecture (RA) for a European service platform. Specification of a set of architectural patterns for non-functional attributes and analysis of how to integrate cloud platforms in the RA.

Highly Scalable Platform for the Construction of Dependable and Ubiquitous Services, *Spanish Ministry of Education and Science (TIN2007-67353-C02)*, 2007–2010, **Roles**: Software Engineer
- Analysis and tests of consistency problems that arise in end-user applications when second-level caches (e.g. Coherence, JBoss cache, etc.) are combined with object persistence mechanisms (e.g. Hibernate)

AUTOMAN: Autonomic Management of Grid-Based Enterprise Services, 2006–2007
- Integration of self-configuration and self-repair properties of autonomic computing in the core of a cloud platform at INRIA (France).

High Performance Distributed Systems, *Community of Madrid (S-0505/TIC/000285)*, 2006–2009, **Roles**: Software Architect, Software Engineer
- Development of a high-available and scalable service for the JOnAS J(2)EE application server. It provided high availability for critical applications deployed in application server clusters, scaling-out the cluster when is overloaded. UPM & BULL signed a pre-agreement to include it in the commercial version. LoC Java (HA&S Service) $\simeq$ 4.000

S4ALL (Services for All), *5th European Framework Programme (IST-2001-37126)*, 2005–2007, **Roles**: Software Architect, Software Engineer
- Development of a high-available service for the JOnAS application server maintained by Bull SAS (France). Available since v.4.8. LoC Java (JOnAS) $\simeq$ 150.000

AUTONOMIC: Autonomic, Dependable and Middleware for Scalable, Distributed, Ubiquitous and Highly Available e-Services, *Spanish Ministry of Education and Science (TIN2004-07474-C02-01)*, 2004–2007, **Roles**: Software Architect, Software Engineer
- Open-source reference implementation of the WS-CAF specification for adding transactions to SOAP Web Services. LoC $\simeq$ 50.000

ADAPT: Middleware for Adaptive and Composable Distributed Components, *EUREKA/ ITEA project (Label 04025)*, 2002–2005, **Roles**: Software Architect, Software Engineer
- Implementation of a transactional-aware replication architecture for stateful EJBs for JBoss. LoC Java $\simeq$ 10.600
- Open-source implementation of the Activity Service specification to add advanced transactions models to J2EE. LoC Java $\simeq$ 10.000

2001–2003 **Lecturer**, *School of CS at Universidad Pontificia de Salamanca (Madrid Campus)*, Spain
Courses on Operating Systems and Programming (C and Pascal)

2000–2001 **Systems Administrator**, *School of CS at Universidad Pontificia de Salamanca (Madrid Campus)*, Spain
Management and maintenance of UNIX/Linux servers and Windows workstations: task automation, security etc.

1999–1999 **Quality Analyst**, *Meta 4 S.A. (now Cegid)*, Madrid, Spain
I tested the database connection modules of Meta4's ERP suite (now Cegid), gaining hands-on experience with multiple DBMSs, including Oracle, Microsoft SQL Server, Informix, and Sybase, as well as JDBC (Java Database Connectivity). My responsibilities included configuring database connections via JDBC, planning and executing tests, analyzing results, and reporting bugs to ensure system reliability and performance.

Academic Research Experience

2003–2011 **Researcher**, *School of CS at Universidad Politécnica de Madrid (UPM)*, Spain
Detailed achievements:

Ph.D. Thesis, *Middleware for High Available and Scalable Multi-Tier and Service-Oriented Architectures*, 2003–2009, Advisors: Prof. Marta Patiño-Martínez and Prof. Ricardo Jiménez-Péris (UPM)
Fields/Topics: *Distributed Systems, Transactional Systems, Scalability, High Availability, SOAs*
- Development of a brand-new approach to provide high availability and scalability to multi-tier architectures by combining snapshot-isolation and an innovative vertical replication approach.

Research Internship, *SARDES Research Group at INRIA Grenoble (France)*, 2007, Advisor: Prof. Sara Bouchenak
- Integration of self-configuration and self-repair properties of autonomic computing in the core of a cloud platform.

Publications in top conferences and journals
- *Elastic SI-Cache: Consistent and Scalable Caching in Multi-Tier Architectures.* In VLDB Journal, 2011 **44 Citations**
- *Scalability Evaluation of the Replication Support of JOnAS, an Industrial J2EE Application Server.* In EDCC Conf., Valencia (Spain), 2010 **10 Citations**
- *A System of Architectural Patterns for Scalable, Consistent and Highly Available Multi-tier Service Oriented Infrastructure.* In Architecting Dependable Systems VI, Springer, 2009 **13 Citations**
- *Consistent and Scalable Cache Replication for Multi-tier J2EE Applications.* In ACM/IFIP/USENIX Middleware Conf., CA (USA), 2007 **43 Citations**
- *WS-Replication: A Framework for Highly Available Web Services.* In ACM WWW Conf., Edinburgh, 2006 **230 Citations**
- *Highly Available Long Running Transactions and Activities for J2EE Applications.* In IEEE ICDCS Conf., Lisbon (Portugal), 2006 **28 Citations**
- *ZenFlow: A Visual Tool for Web Service Composition.* In IEEE VL/HCC Conf., Dallas (USA), 2005 **53 Citations**

Technical Skills

O.O. design	Design patterns, agile techniques, UML
Lang. & Fmwk.	Python, Pytorch/TF/Scikit, HFT, Java, Go, Rust, Ray, Pandas, Dask, /ldots
CI/CD tools and services	VCS (Git/Github/GitLab)
Cloud & Virt.	AWS, GCloud, Docker, Kubernetes
DBMS	Relational and NoSQL DBs
O.S. admin.	UNIX flavors, Android, iOS, Windows
Misc.	VSCode, IntelliJ, Vi, Emacs, UNIX shell scripting, LaTeX, HTML, basic Javascript, etc.

Education

2011 **Certificate of Training, Advanced Scala**, *Typesafe*, Switzerland

2011 **Certificate of Training, Scala**, *Typesafe*, Switzerland

2011 **Course in Business Administration and Economics (268 hours)**, *Funded by Community of Madrid*, Spain

2003-2009 **Ph.D. in CS**, *School of CS at Universidad Politécnica de Madrid (UPM)*, Spain

2004 **Postgraduate Certificate in Education**, *Educational Sciences Institute at Universidad Complutense de Madrid (UCM)*, Spain

1994–2001 **B.Eng & M.Sc. in CS**, *School of CS at Universidad Pontificia de Salamanca (Madrid Campus)*, Spain

Communication Skills

Delivered talks and presentations at both international and national conferences, effectively communicating complex ideas to diverse audiences. Presented research findings and project updates in internal meetings within corporate environments. Additionally, demonstrated teaching and mentoring skills by conducting undergraduate and Ph.D.-level courses at the university, fostering understanding and engagement among students.

Languages

Spanish *Mother tongue*

English *Fluent (written and spoken)*

French *Intermediate proficiency (listening, reading, & speaking)*

Catalan *Intermediate proficiency (listening, reading, & speaking)*

Other Activities Related to Computer Science

Committer in the Apache Software Foundation

- *Apache Omid project, a high-performant and scalable Transaction Manager for HBase*

Reviewer in International Academic Conferences

- *IEEE International Symposium on Reliable Distributed Systems (SRDS)*, 2008 and 2009
- *IEEE International conference in Distributed Computing Systems (ICDCS)*, 2009
- *International Conference on Parallel and Distributed Computing (Europar)*, 2007 and 2010
- *ACM Symposium on Applied Computing (SAC)*, 2008 and 2010
- *International Conference on Service-Oriented Computing (ICSOC)*, 2009
- *EDBTs International Workshop on Data Management in Peer-to-peer Systems (DAMAP)*, 2009
- *International Workshop on Assurance in Distributed Systems and Networks (ADSN)*, 2010

Speaker/Attendee in Technical Conferences

- *Neurips 2024 conference*, Dec 10th-15th. 2024, Vancouver, BC (Canada)
- *ICML 2024 conference*, Jul 21st-27th. 2023, Vienna (Austria)
- *MLSys 2024 conference*, May 13th-16th. 2023, Santa Clara, CA (USA)
- *Neurips 2023 conference*, Dec 10th. 2023, New Orleans, LA (USA)
- *Amazon Re:Invent 2023*, Nov 27th-30th. 2023, Las Vegas, NV (USA)
- *Ray Summit 2023*, Sep 18th-20th. 2023, San Francisco, CA (USA)
- *ICML 2023 conference*, Jul 23rd-29th. 2023, Honolulu, Hawaii (USA)
- *MLSys 2023 conference*, Jun 4th-8th. 2023, Miami, FL (USA)
- *Ray Summit 2022*, Aug 23th-24th. 2022, San Francisco, CA (USA)
- *MLSys 2022 conference*, Aug 31th-Sept 3rd. 2022, Santa Clara, CA (USA)
- *ACL 2019 conference (Speaker)*, Jul 28th-Aug 2nd. 2019, Florence (Italy)
- *@Scale conference*, 31st Aug. 2016, San Jose, CA (USA)
- *Hadoop Summit (Speaker)*, 28th-30th Jun. 2016, San Jose, CA (USA)
- *@Scale conference*, 14th Sep. 2015, San Jose, CA (USA)
- *Hadoop Summit Europe*, 15th-16th Apr. 2015, Brussels (Belgium)
- *Hadoop Summit Europe*, 2nd-3rd Apr. 2014, Amsterdam (Netherlands)
- *NoSQL Matters Conf.*, 21st-22nd Nov. 2014, Barcelona (Spain)
- *NoSQL Matters Conf.*, 29th-30th Nov. 2013, Barcelona (Spain)
- *NoSQL Matters Conf.*, 6th Oct. 2012, Barcelona (Spain)
- *IEEE International Symposium on Reliable Distributed Systems (SRDS)*, 4-7th Oct. 2011, Madrid (Spain)
- *World Wide Web Conference*, 20-24th Apr. 2009, Madrid (Spain)
- *Spanish Conference on Concurrency and Distributed Systems*, 13-16th Sep. 2005, Granada (Spain)
- *ObjectWebCon'05*, 17-20th Jan. 2005, Lyon (France)
- *ObjectWeb's Workshop on Transactions*, 23-24th Feb. 2004, Grenoble (France)
- *ObjectWeb's Architecture Meeting*, 13-15th Jan. 2004, Sevilla (Spain)

Member

- *ObjectWeb Consortium*, http://www.ow2.org
- *Java Community Process program (JCP)*, http://jcp.org

Hobbies and Interests

Beyond my core professional focus, but some way related to it, I have a strong interest in the intersection of fields such as neuroscience, psychology, decision-making, cognitive sciences, learning techniques, behavioral economics, and philosophy, as they offer valuable insights into human behavior, intelligence, and how the brain works (and why!) I am deeply fascinated by the mechanisms of human thought, learning, and behavior, and how these insights can inform and inspire advancements in AI and our daily lives and wellbeing in general.

In my personal life, I enjoy staying active through activities like running, mountain biking, yoga, cold-plunging, and playing tennis, which not only keep me physically fit but also provide a sense of balance and focus.