

JOÃO FONSECA

jfonseca.eu

joaofonseca@inesc-id.pt ♦ joaofonseca@nyu.edu

Google Scholar (ID: WcekujkAAAAJ)

POSITIONS

Assistant Professor, Faculty of Computer Science, University of Lisbon Faculdade de Ciências - Universidade de Lisboa	Lisbon, Portugal Sep 2026 – Present
Visiting Scholar, Center for Responsible AI, New York University <i>Advisor: Julia Stoyanovich</i>	Remote Oct 2025 – Present
Postdoctoral Fellow, Distributed, Parallel and Secure Systems, INESC-ID <i>Advisors: Paolo Romano & Rodrigo Rodrigues</i>	Lisbon, Portugal Nov 2025 – Aug 2026
Postdoctoral Fellow, Center for Responsible AI, New York University <i>Advisor: Julia Stoyanovich</i>	New York, NY, USA Oct 2024 – Sep 2025
Invited Assistant Professor, NOVA University Lisbon <i>NOVA Information Management School</i>	Lisbon, Portugal Sep 2023 – Sep 2024
Visiting Scholar, Center for Responsible AI, New York University <i>Advisor: Julia Stoyanovich</i>	Remote Jul 2023 – Sep 2024
Research Fellow, NOVA University Lisbon <i>Advisor: Fernando Bação NOVA Information Management School</i>	Lisbon, Portugal Apr 2019 – Sep 2020

EDUCATION

NOVA University Lisbon Ph.D. in Information Management (NOVA Information Management School) M.S. in Information Management (NOVA Information Management School) M.S. in Management (NOVA School of Business and Economics) B.S. in Economics (NOVA School of Business and Economics)	Lisbon, Portugal Sep 2020 – Oct 2023 Sep 2017 – Jan 2019 Sep 2016 – Jan 2019 Sep 2013 – Jun 2016
New York University Visiting PhD student / Research Intern (Tandon School of Engineering)	New York, NY, USA Mar 2023 – Jun 2023
Universidad de Granada Exchange Semester (Facultad de Ciencias Económicas y Empresariales)	Granada, Spain Sep 2015 – Feb 2016

TEACHING

Computation II: Algorithms and Data Structures <i>Course coordinator NOVA Information Management School</i>	NOVA University Lisbon Feb 2024 – Jul 2024
Computation III: Object-oriented programming <i>Course coordinator NOVA Information Management School</i>	NOVA University Lisbon Sep 2023 – Jan 2024
Data Mining <i>Teaching Assistant NOVA Information Management School</i>	NOVA University Lisbon Sep 2019 – Jan 2024
Business Cases with Data Science <i>Teaching Assistant NOVA Information Management School</i>	NOVA University Lisbon Feb 2022 – Jul 2022
Web Analytics <i>Teaching Assistant NOVA School of Business and Economics</i>	NOVA University Lisbon Feb 2020 – Jul 2020

SUPERVISIONS

PhD Thesis	- Mafalda Sá Velho, “A New Paradigm in Explainable AI: Combining Large Language Models and XAI for Narrative Explanations in Black-Box Models”	2025 — PRES.
Masters Thesis	- Duarte Girão, “What must a player improve to increase its overall ranking?”	2024
	- Reinaldo dos Santos Barros, “Improving relationship channels in the pharmaceutical industry with machine learning”	2024

PUBLICATIONS

Asterisks denote equal contribution or alphabetical ordering. Also see Google Scholar.

Notable venues: EMNLP 2025 (Core A*), VLDB 2025 (Core A*), SIGMOD 2024 (Core A*), Journal of Big Data (Q1), Expert Systems with Applications (Q1), International Journal of Intelligent Systems (Q1)

- [19] **J. Fonseca**, R. Rodrigues, and P. Romano, “Mixture-of-expert blocks contain strong hallucination detection signals,” *arXiv preprint arXiv:2608.17687 (UNDER SUBMISSION)*, 2026. [Online]. Available: <https://doi.org/10.48550/arXiv.2608.17687>
- [18] **J. Fonseca** and J. Stoyanovich, “Explainerpfn: Towards tabular foundation models for model-free zero-shot feature importance estimations,” *arXiv preprint arXiv:2601.23068 (UNDER SUBMISSION)*, 2026. [Online]. Available: <https://doi.org/10.48550/arXiv.2601.23068>
- [17] S. Lee, **J. Fonseca**, and J. Stoyanovich, “Relshap: Relationally consistent shapley explanations,” *arXiv preprint arXiv:2608.11508 (UNDER SUBMISSION)*, 2026. [Online]. Available: <https://doi.org/10.48550/arXiv.2608.11508>
- [16] A. Bell*, **J. Fonseca***, C. Abrate, F. Bonchi, and J. Stoyanovich, “How much effort is enough? fairness in algorithmic recourse through the lens of substantive equality of opportunity,” in *Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, ser. EAAMO ’25, New York, NY, USA: Association for Computing Machinery, 2025, pp. 170–184. [Online]. Available: <https://doi.org/10.1145/3757887.3763014>
- [15] **J. Fonseca***, A. Bell*, and J. Stoyanovich, “SAFENUDDGE: Safeguarding large language models in real-time with tunable safety-performance trade-offs,” in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP ’25, Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 19 955–19 969. [Online]. Available: <https://doi.org/10.18653/v1/2025.emnlp-main.1010>
- [14] H. Hwang, A. Bell, **J. Fonseca**, V. Pliatsika, J. Stoyanovich, and S. E. Whang, “SHAP-based explanations are sensitive to feature representation,” in *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT ’25, New York, NY, USA: Association for Computing Machinery, 2025, pp. 1588–1601. [Online]. Available: <https://doi.org/10.1145/3715275.3732105>
- [13] V. Pliatsika*, **J. Fonseca***, K. Akhynko, I. Shevchenko, and J. Stoyanovich, “Sharp: Explaining rankings and preferences with shapley values,” *Proc. VLDB Endow.*, vol. 18, no. 11, pp. 4131–4143, Jul. 2025. [Online]. Available: <https://doi.org/10.14778/3749646.3749682>
- [12] A. Bell*, **J. Fonseca***, and J. Stoyanovich, “The game of recourse: Simulating algorithmic recourse over time to improve its reliability and fairness,” in *Companion of the 2024 International Conference on Management of Data*, ser. SIGMOD ’24, Santiago, Chile: Association for Computing Machinery, 2024, pp. 464–467. [Online]. Available: <https://doi.org/10.1145/3626246.3654742>
- [11] **J. Fonseca**, “The role of synthetic data in improving supervised learning methods: The case of land use/land cover classification,” Ph.D. dissertation, Universidade NOVA de Lisboa (Portugal), 2023.
- [10] **J. Fonseca** and F. Bacao, “Geometric smote for imbalanced datasets with nominal and continuous features,” *Expert Systems with Applications*, vol. 234, p. 121 053, 2023. [Online]. Available: <https://doi.org/10.1016/j.eswa.2023.121053>
- [9] **J. Fonseca** and F. Bacao, “Improving active learning performance through the use of data augmentation,” *International Journal of Intelligent Systems*, vol. 2023, no. 1, p. 7 941 878, 2023. [Online]. Available: <https://doi.org/10.1155/2023/7941878>
- [8] **J. Fonseca** and F. Bacao, “Tabular and latent space synthetic data generation: A literature review,” *Journal of Big Data*, vol. 10, no. 1, p. 115, 2023. [Online]. Available: <https://doi.org/10.1186/s40537-023-00792-7>

- [7] **J. Fonseca***, A. Bell*, C. Abrate, F. Bonchi, and J. Stoyanovich, “Setting the right expectations: Algorithmic recourse over time,” in *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, ser. EAAMO ’23, Boston, MA, USA: Association for Computing Machinery, 2023. [Online]. Available: <https://doi.org/10.1145/3617694.3623251>
- [6] **J. Fonseca** and F. Bacao, “Research trends and applications of data augmentation algorithms,” *arXiv preprint arXiv:2207.08817*, 2022. [Online]. Available: <https://doi.org/10.48550/arXiv.2207.08817>
- [5] **J. Fonseca**, G. Douzas, and F. Bacao, “Improving imbalanced land cover classification with k-means smote: Detecting and oversampling distinctive minority spectral signatures,” *Information*, vol. 12, no. 7, 2021. [Online]. Available: <https://doi.org/10.3390/info12070266>
- [4] **J. Fonseca**, G. Douzas, and F. Bacao, “Increasing the effectiveness of active learning: Introducing artificial data generation in active learning for land use/land cover classification,” *Remote Sensing*, vol. 13, no. 13, 2021. [Online]. Available: <https://doi.org/10.3390/rs13132619>
- [3] A. Crayton*, **J. Fonseca***, K. Mehra*, M. Ng*, J. Ross*, M. Sandoval-Castaneda*, and R. von Gnechten*, “Narratives and needs: Analyzing experiences of cyclone amphan using twitter discourse,” in *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*, 2020. [Online]. Available: <https://www.climatechange.ai/papers/neurips2020/42>
- [2] G. Douzas, F. Bacao, **J. Fonseca***, and M. Khudinyan*, “Imbalanced learning in land cover classification: Improving minority classes’ prediction accuracy using the geometric smote algorithm,” *Remote Sensing*, vol. 11, no. 24, 2019. [Online]. Available: <https://doi.org/10.3390/rs11243040>
- [1] **J. Fonseca**, “Harnessing big data to inform tourism destination management organizations,” M.S. thesis, Universidade NOVA de Lisboa (Portugal), 2018.

HONORS AND AWARDS

Cornell Tech – Runway Startup Postdoc Program (Rejected Offer)	2025
Catalyst NYC Accelerator – Pre-Seed Startup Accelerator	2025
Tech Venture Workshop – Non-Dilutive Startup Funding	2024
Best AI Track Paper Award – ACM EAAMO’23 Conference	2023
MIT Portugal Graduate Research Fellowship	2020
DSSG Summit 2020 Challenge – Hackathon Winner	2020

RESEARCH PROJECTS

Hallucination Detection in Large Language Models <i>European Union’s Horizon Europe Grant Agreement No 101189689</i>	Nov 2025 – Present
Project developed with Professors Paolo Romano, PhD, and Rodrigo Rodrigues, PhD. Developing a framework to detect hallucinations in Large Language Models (LLMs) leveraging internal signals produced by Mixture-of-Experts (MoE) architectures.	
Machine Learning Explainability Frameworks <i>National Science Foundation (USA) NSF grants 2326193, 2312930</i>	Mar 2023 – Present
- Developing DiffusionExplainerPFN, a Diffusion-based approach to address the limitations found while developing ExplainerPFN. - Contributed to the development of RelSHAP, a framework that produces relationally consistent Shapley explanations. - Proposed ExplainerPFN, a model-agnostic framework for interpreting model predictions using a probabilistic approach. - Proposed ShaRP (Shapley for Rankings and Preferences), a framework that explains the contributions of features to different aspects of a ranked outcome, and is based on Shapley values. Currently counts with over 5,000 downloads.	

ML-Research — An Open Source Library for Machine Learning Research

Sep 2020 – Present

MIT Portugal | FCT grant SFRH/BD/151473/2021

- ML-Research contains the software implementation of most algorithms used or developed in my research. It contains Active Learning and Oversampling implementations (among others), as well as Datasets and several utilities to conduct research. Currently counts with **over 158,000 downloads**.

Auditing and Safeguarding Large Language Models

National Science Foundation (USA) | NSF grants 2520637, 2326193, 2312930

Sep 2024 – May 2026

- Developed Safe-Lens, an interactive platform for visualizing Large Language Model inference in real time, with token-level probability analysis and automated content moderation via SafeNudge. Currently being used for teaching purposes at NYU. **Publicly available at <https://safe-lens.users.hsrn.nyu.edu/>**
- Developed SafeNudge, a real-time safeguarding method against red teaming Large Language Models.
- Developed Output-Scouting, an open-source library designed to search for dangerous/unsafe outputs within Large Language Models' possible output space. This was an exploratory project that led to SafeNudge.

Game of Recourse — Multi-Agent Algorithmic Recourse Over Time

Mar 2023 – Jul 2025

MIT Portugal | FCT grant SFRH/BD/151473/2021

- Developed a multi-agent framework for Algorithmic Recourse, which allows agents to interact with each other and the environment to compete for a scarce, favorable outcome.
- Recourse-Game contains the software implementation of the Multi-agent Algorithmic Recourse framework proposed in the paper “Setting the Right Expectations: Algorithmic Recourse Over Time.”

MapIntel — Interactive Visual Analytics Platform for Competitive Intelligence

Fundação para a Ciência e Tecnologia | FCT grant DSAIPA/DS/0116/2019

Feb 2021 – Aug 2021

- Project developed with Professor Fernando Bação, PhD and Georgios Douzas, PhD. This work focused on the development of an interactive visual analytics platform for competitive intelligence.
- Usage of various Natural Language Processing and visualization techniques.

IPSTERS — IPSentinel Terrestrial Enhanced Recognition System

Fundação para a Ciência e Tecnologia | FCT grant DSAIPA/AI/0100/2018

Apr 2019 – Jan 2021

- Worked with Professor Fernando Bação, PhD. Exploration of applications and limitations of artificial intelligence algorithms to perform automatic updates of Land Use/Land Cover maps for the Portuguese mainland using satellite imagery.
- Exploration of new data filtering/anomaly detection algorithms, as well as dimensionality reduction techniques and different types of classification algorithms, including ensemble methods and deep learning classifiers.

Amphan — Analyzing Experiences of Extreme Weather Events using Online Data

DSSG Solve for Good, Data Science for Social Good Foundation | Volunteer Research Project Jul 2020 – Nov 2020

- Integrated a team of four data scientists to research and develop a method to analyze social media discourse around natural disasters in developing countries in partnership with the International Water Management Institute (IWMI).
- Collaborative effort to develop data collection, preprocessing, feature extraction and analysis methods. Most of the tasks done were centered around the exploration of state-of-the-art Natural Language Processing algorithms.

Research + Master's Thesis (Final Grade: 19 / 20)

NOVA School of Business and Economics, NOVA University Lisbon

Oct 2017 – Mar 2019

- Worked with Professor Leid Zejnilovic, PhD. Used digital traces created by tourists to understand and improve the efficiency of tourism flows in Portugal, particularly specifically in the region of Lisbon.
- Studied the potential of Big data to inform destination management organizations using three sources of Big data: Telecom, Social media and Airbnb data.
- Developed a Social Media Crawler to fetch, store and visualize data of three major social media channels (Facebook, Twitter and Instagram).

SELECTED INVITED TALKS

NYC Responsible AI Research Day, “ExplainerPFN: Towards Tabular Foundation Models for Model-Free Zero-Shot Feature Importance Estimations”, New York University, New York AI Exchange	Mar 27, 2026
Promoting Responsible Innovation in Artificial Intelligence for Peace and Security exhibit, “Why do Large Language Models Hallucinate?”, United Nations, Office for Disarmament Affairs	May 7, 2025
AI Congressional Demo Speaker, U.S. Capitol Hill, Center for AI Policy	Feb 24, 2025
DSSG Summit 2020 – Presentation of the winning submission of the community challenge: “Keyboard layout optimization for ALS patients competition”	Oct 21, 2020
Solve for Good - DataFest 2020, “Amphan: Analyzing experiences of extreme weather events using online data”	Sep 17, 2020
Data Science for Hospitality and Tourism conference, “Unleashing the power of big data”, Panel member, Rectorate of NOVA University Lisbon	Dec 4, 2018
NOVA School of Business and Economics’ Inauguration Ceremony, “Data-driven planning for sustainable tourism in Portugal”	Sep 29, 2018

COMMUNITY SERVICE

- **Reviewer:** ICLR (2026, 2025), FAccT 2026, NeurIPS 2026
- **Proceedings Chair:** FAccT 2026

LANGUAGES

Fluent in Portuguese, English and Spanish; Basic French.