

Pedro Rodriguez | Research Scientist

✉ pedro@snowgeek.org • 🌐 pedrorodriguez.io

Experience

University of Maryland at College Park

Research Assistant to Jordan Boyd-Graber, PhD Candidate

Machine Learning, Deep Learning, NLP

August 2015 - Present

- Research experience include machine learning, deep learning, natural language processing, and question answering.
- Created an AI for a trivia game (Quiz Bowl) that beats the best human players, and writing journal paper publication
- Team Captain and founding member of the Colorado Data Science Team: leadership and technical role promoting data science education through applied experience and creating synergistic industry partnerships at CU Boulder.

Riot Games

Data Science Research Intern

Machine Learning, Natural Language Processing, Bayesian Modeling

May 2017 - August 2017

- Improved NLP model effectiveness for automatically sorting and/or answering 12,000 player support tickets per day.
- Created data labeling process for player support tickets to increase effectiveness and enable incremental future improvement.
- Used Machine Learning ablation studies, Bayesian modeling and Monte Carlo simulations to understand behavior of new sportsmanlike play system called Honor 2.0. This also included querying over 56 million records with Spark SQL.

Oracle Data Cloud/Datalogix

Data Science Research Intern

Data Engineering, Machine Learning, DevOps

May 2016 - August 2016

- Architected and implemented ETL pipeline with Apache Spark/Airflow which ingests 200GB/day of advertising data.
- Improved throughput from AWS S3 to Apache Spark by 40-60x which resulted in substantial cost savings.
- Created data science workflows using Packer, Terraform, and Zeppelin which greatly increased productivity coworkers.

Zillow Group/Trulia

Data Scientist

Crowdsourcing, DevOps, Machine Learning

January 2015 - July 2015

- Designed and iteratively improved crowdsourcing experiments to produce rank order for machine learning generated photo albums. In doing so wrote a Amazon Mechanical Turk library which managed approximately 20,000 tasks.
- Architected and implemented infrastructure for internal data science APIs built on Docker, Swagger, Jenkins, and predominantly Django as a web framework. This code is now called on 100% of visits to trulia.com.
- Ran educational workshops within the company to facilitate the use and learning of Apache Spark and Docker.

UC Berkeley AMPLab

Undergraduate Research Assistant

Distributed Machine Learning

Fall 2014

- Wrote optimized implementation of Gibbs sampling Latent Dirichlet Allocation for topic modeling on Apache Spark.

Boise State University Geophysics: Cryosphere Scientist, Field Researcher, May 2013-October 2014

UC Berkeley Center for Time Domain Informatics: Undergraduate Researcher in Astrophysics, 2012

Avalanche Safety Educator: Volunteer for National Ski Patrol, Professionally in South America, 2010-Present

Projects and Code

Qanta: An AI that plays quiz bowl with deep learning on PyTorch, Apache Spark, Vowpal Wabbit, Terraform, Packer, Amazon Web Services, and more to create running a robust state-of-the-art system

Apache Spark: Contributed to implementation of topic modeling algorithm (LDA), and expansion of SparkSQL built in functions. Code is in MLlib, Intel Analytics Topic Model package, and Spark 1.5

PyFunctional: Author of popular python package exposing a collections API borrowed from Scala/Spark

Code: github.com/EntilZha, bitbucket.org/pedrorodriguez, pedrorodriguez.io

Languages: Python, SQL, Scala, HTML/CSS/Javascript, Tex, Rust, Go, C/C++, Java, Ruby

Education

University of Maryland at College Park, Computer Science

PhD Candidate specializing in Machine Learning, Deep learning, and NLP

Expected Spring 2020

University of California at Berkeley, Computer Science

Bachelor of Arts in Computer Science with coursework in Physics and Astrophysics

2010 - 2014