# José Pablo Cambronero

## Overview and Research Interests

(As of March 2, 2020) I am a fourth year student in the EECS PhD program at MIT. I am currently working at the intersection of software engineering and machine learning, and am interested in exploring applied machine learning research.

## Education

### Academic Qualifications

2016-TBD **PhD EECS Candidate**
*Massachusetts Institute of Technology*, Cambridge, MA.

2013-2016 **Masters in Computer Science**
*New York University: Courant Institute of Mathematical Sciences*, NY, NY.
*GPA: 3.89, MS Research/Thesis Fellowship Award Fall 2015, funding work on A2Q (an order-aware optimizing query compiler for AQuery)*

2007-2011 **Bachelor of Arts in Economics and Minor in German Studies**
*University of Pennsylvania*, Philadelphia, PA.
*GPA: 3.93, Phi Beta Kappa, Summa Cum Laude, Dean's List (08, 09, 10)*

## Relevant Coursework

- MIT: Computer Architecture, Theory of Computation, Database Systems, Machine Learning
- NYU: Compiler Construction, Natural Language Processing, Speech Recognition, Programming Languages, Rigorous Software Development (an introduction to formal methods), Principles of Software Security

## Academic Work Experience

2015 – 2016 **Graduate Course in Compiler Construction** *Grader*, NYU.

Fall 2014 **Graduate Course in Programming Languages** *Teaching Assistant*, NYU.

## Industry Work Experience

Fall 2018 **Part-Time Research Collaborator** *Big Code Team*, Facebook, Remote.

- Applying deep learning to identify and highlight core code functionality.

Summer 2018 **Intern** *Software Engineering*, Facebook, Boston.

- Worked with the Big Code team on applications of neural networks to code search
- Implemented different models, carried out evaluation, and collaborated on paper writing
- Study compared techniques to state-of-the-art, showing our simpler networks are competitive and in some cases out perform more complex architectures
- We identified key challenges to neural code search across corpora
- Work presented at FSE 2019

☎ *215-900-5308* • ✉ *jcamsan@mit.edu* • 🌐 *www.josecambronero.com*
*www.github.com/josepablocam*

Summer 2015 **Intern** *Data Science*, Cloudera, San Francisco.

- Contributed multiple statistical tests and classical model implementations to a time series library for Spark (Github: Link)
- Contributed a distributed implementation of Kolmogorov-Smirnov test to Spark-MLlib (Github: Link)
- Wrote blog posts detailing technical contributions and use of time series library. (Blog: Link)

2011 – 2014 **Full-Time Securitized Credit Research Associate** *Non-Agency Mortgages and US Housing*, Morgan Stanley, New York.

- Developed group analytics infrastructure to drive independence from tools built/maintained by quant team
- Learned q programming language independently, quickly became productive in the language, frequently helping others with technical q questions and eventually helping in the review process of the latest *Q for Mortals* (Borror 2016) book
- Introduced R development into the group and wrote base libraries for group
- Led development of various research reports and investing themes

Summer 2010 **Richard B. Fisher Scholar** *Fixed Income Generalist Sales and Fixed Income Credit Strategy*, Morgan Stanley, New York.

Summer 2009 **Douglas Paul Scholar** *Investment Banking and Alternative Investments*, Morgan Stanley, New York.

## Ongoing Research

- **Electronic Health Records for Predictive Diagnoses** We apply deep learning and classic machine learning techniques for early prediction of pancreatic cancer diagnoses. Joint work with Limor Appelbaum (Beth Israel Deaconess Medical Center) and Martin Rinard.

## Past Research

- **Automating construction of machine learning pipelines based on existing programs**: We learn to generate programs implementing classical machine learning pipelines (preprocessing, model fitting, evaluating) by applying dynamic analysis to existing, crowd-sourced, programs. We build a language model for pipelines and use this to guide a search over component choices. Joint work with Martin Rinard. (Publication [4]).
- **Active Learning for Software Engineering**: We present multiple systems that use active learning to infer and re-generate software applications. We show that modularity provides an opportunity to scale these techniques to real-world applications. We characterize the broader paradigm, along with open research questions. Joint work with Thurston Dang, Nikos Vasilakis, Jiasi Shen, Jerry Wu and Martin Rinard. (Publication [2]).
- **User study evaluating the effectiveness of automated program repair**: We designed and carried out a study where a group of MIT graduate students was tasked with repairing open source bugs. We evaluated the potential benefits in terms of bugs solved when given access to an existing state-of-the-art program repair tool. Joint work with Jiasi Shen, Jürgen Cito, Elena Glassman and Martin Rinard. (Publication [5]).
- **ImputeDB**: A database query optimizer for replacing missing values (imputation). ImputeDB incorporates the placement of imputation operators into planning and allows users to balance query quality and execution speed. We show that our technique provides orders-of-magnitude speed up over the prevailing approach and introduce little error in most cases. Joint work with John Feser, Micah Smith, and Samuel Madden. (Publication [1])(Github: Link).
- **DaltonQuant**: A novel image quantization technique tailored to individuals with color vision deficiencies. We build user-specific color confusion quantification functions using a large dataset collected through an iOS game about color, and use this in a multi-objective constrained optimization formulation of color quantization. Our technique reduces file sizes by 22%-29% over the state-of-the-art techniques. Joint work with Phillip Stanley-Marbell and Martin Rinard. (Github: Link).

- **A2Q**: A compiler with pattern-based optimizations targeting time series queries. Written in Scala and based on existing research by Alberto Lerner and Dennis Shasha. Joint work with Dennis Shasha. (Github: Link).

## Publications

[1] Jose Cambronero, John Feser, Micah Smith, and Samuel Madden. Query optimization for dynamic imputation. *PVLDB*, 10(11):1310–1321, 2017.

[2] José P. Cambronero, Thurston H.Y. Dang, Nikos Vasilakis, Jiasi Shen, Jerry Wu, and Martin Rinard. Active Learning for Software Engineering. In *SPLASH Onward!*, 2019.

[3] José P. Cambronero, Hongyu Li, Seohyun Kim, Koushik Sen, and Satish Chandra. When Deep Learning Met Code Search. In *FSE (Industry Track)*, 2019.

[4] José P. Cambronero and Martin Rinard. AL: Autogenerating Supervised Learning Programs. In *SPLASH OOPSLA*, 2019.

[5] José P. Cambronero, Jiasi Shen, Jürgen Cito, Elena Glassman, and Martin Rinard. Characterizing Developer Use of Automatically Generated Patches. In *VL/HCC (Short Paper)*, 2019.

## Language skills

- **Programming Languages:** Proficient in: Python, Java, C, q, R, Scala.
- **Natural Languages:** Native fluency in English and Spanish. Working proficiency in German.

## Service

- **Artifact Evaluation Committee CAV 2020**
- **Artifact Evaluation Committee PPoPP 2018**
- **MIT PL Offsite 2017**: I co-organized, with Ivan Kuraj, the MIT Programming Languages offsite 2017. The event is meant to foster dialogue and ideas among members of the MIT PL community and neighboring institutions.
- **MIT Admitted Students' Visit Weekend Diversity Panel (2017, 2019, 2020)**: I co-organized a diversity panel aimed to provide a venue for prospective students to ask any questions they might have about diversity at MIT and how we are working towards improving our community.
- **CSAIL Student Committee (2017 - 2020)**: I serve as Treasurer on the CSAIL Student Committee. I manage the group's budget and contribute with the organization of social events, such as a weekly event featuring baked goods and socializing among graduate students in CSAIL.