

Overview and Research Interests

(As of October 18, 2018) I am a third year student in the EECS PhD program at MIT. My research interests center around improving developer tools (particularly for areas such as databases and data science) by learning from existing codebases.

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.*
(Ongoing)
 - Applying code embeddings to identify core method functionality
 - Goal is to provide information to facilitate code review and code search by highlighting the core functionality of a code fragment
- 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 currently under submission

44 Parker Hill Ave, Boston MA 02120

☎ 215-900-5308 • ✉ jcamsan@mit.edu • 🌐 www.josecambronerero.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

- **Code Search for Data Analysis with Execution Data:** We are developing a code search tool that targets data analysis. Rather than index a large, potentially unrelated collection of programs, it creates an index associated with a particular dataset by only considering programs that were written to work with that dataset. Users of the search tool provide a natural language query for their task, along with type-based predicates that a relevant API call should satisfy. Our system uses these two inputs to search execution traces and documentation for potential key APIs. We then identify call sites for these APIs and perform program slicing to extract the relevant code fragment. Joint work with Martin Rinard and Raul Castro Fernandez.
- **Automating construction of machine learning pipelines based on existing programs:** We collected a large amount of executable programs that solve supervised learning problems. We analyzed their execution and model the choice of pipeline operators based on existing code and data characteristics. This model is then used to generate new pipelines for new datasets by enumerating possible operator choices and pruning partial programs based on the likelihood that a developer would write a similar pipeline. (In submission) Joint work with Martin Rinard

Past Research

- **User study evaluating the effectiveness of automated program repair:** We designed and executed 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 and Martin Rinard.
- **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 (Publication [1])(Github: [Link](#)). Joint work with John Feser, Micah Smith, and Samuel Madden
- **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

44 Parker Hill Ave, Boston MA 02120

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

work with Phillip Stanley-Marbell and Martin Rinard

- **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. (Github: [Link](#)) Joint work with Dennis Shasha

Publications

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

Technical skills

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

Service

- **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.
- **2017 MIT Admitted Students' Visit Weekend Diversity Panel**: I co-organized, with Candace Ross, the first diversity panel during admitted students' visit weekend at MIT. The 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**: 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.

44 Parker Hill Ave, Boston MA 02120

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