Overview and Research Interests

(As of May 19, 2021) I recently defended my PhD thesis, supervised by Martin Rinard, in the EECS department at MIT. I work at the intersection of software engineering and machine learning, and am broadly interested in applying machine learning to improve developer and data analyst tools. I will be joining the PROSE team at Microsoft as a researcher starting the summer of 2021.

Education

Academic Qualifications

2016-2021 **PhD in Computer Science**
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

Summer 2020 **Intern** *Facebook AI Research*, Facebook, Remote (due to COVID-19).
- Worked with the SysML team on a novel tensor compiler
- Contributed C++ code to a just-in-time compiler for high-performance tensor operations
- Wrote a C++ translator from loop order/shape specifications to Halide programs
- Implemented benchmarking for key tensor operations such as matrix multiplication, convolutions, and matrix/vector products
- Wrote a Python translator from loop order/shape specifications to TVM programs

Fall 2018 **Part-Time Research Collaborator** *Big Code Team*, Facebook, Remote.
- Applying deep learning to identify and highlight core code functionality

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

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

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

- **Mining nearby transformations to improve machine learning pipelines** We frame improving machine learning pipeline's performance as a program repair problem. Our system extracts typed edit rules associated with pipeline performance improvement, summarizes these, and applies them to produce improved pipelines. Our repair system improves over random mutation, a popular alternative for "short-step" edits. Joint work with Jürgen Cito, Micah Smith, and Martin Rinard. (Under submission [1])
- **Dataset sampling for AutoML** We investigate the impact of dataset downsampling on a genetic programming (GP) based AutoML tool. Dataset downsampling presents an appealingly simple solution to the challenge of scaling AutoML search techniques to increasingly large datasets. Our results show that downsampling allows the GP search to explore substantially more pipelines, resulting in improved performance when compared to carrying out search on the full dataset. Joint work with Fatjon Zogaj (equal contrib.), Martin Rinard, and Jürgen Cito. (Under submission [2])
- **Mining wrangling functions from Python programs** We use dynamic program analysis to extract data preparation steps from a collection of Python programs written to analyze the same dataset. These data transformations are modularized into Python functions, which we term wrangling functions. We store these functions in a database, which a new analyst can query to obtain transformations and integrate these into their own computational notebook, reusing prior analysts efforts. Joint work with Raul Castro Fernandez and Martin Rinard. (Under submission [3])
- **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. (Publication [4])

Past Research

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

- **Automating construction of AutoML search spaces**: AutoML tools can facilitate the process of automatically producing high performance machine learning pipelines. However, they do not allow users to easily express preferences for pipelines in the search space. We propose a model where users provide preferences as a set of API components. This set is a form of *weak specification*, which our system (AMS) automatically augments with additional alternative components, complementary components, and a defined set of hyperparameters and values for tuning. To automatically perform this extension AMS mines API documentation and existing code examples. Joint work with Jürgen Cito and Martin Rinard. (Publication [5]).
- **Pancreatic Cancer Risk Modeling**: We developed a diagnoses-based risk prediction model for early detection of pancreatic cancer (PDAC). We compare multiple model alternatives, and carry out internal and external validation. Our results show that our diagnoses-based model can outperform a sparse model based on clinically-known indicators and provides an initial and promising direction for EHR-based PDAC screening. Joint work with Limor Appelbaum (co-lead author), Martin Rinard, and others at Beth Israel Deaconess Medical Center, Brigham and Women's Hospital, Harvard University, Dana Farber Cancer Institute, and MIT. (Publication [4]).
- **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 [6]).
- **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 [7]).
- **Deep learning for code search** We investigate the use of deep learning for code search based on natural language queries. We propose a simplified neural architecture and compare it to existing neural-based code search techniques. Our results show that a simple model can outperform more sophisticated, sequenced-based networks. We carry out additional analysis and evaluation to investigate performance in different corpora. Work done while an intern at Facebook during summer of 2018. Joint work with Hongyu Li, Seohyun Kim, Koushik Sen, and Satish Chandra. (Publication [8]).
- **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 [9]).
- **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 [10])(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] José P. Cambronero, Micah Smith, Jürgen Cito, and Martin Rinard. Learning Repair Rules for Machine Learning Pipelines from AutoML Search Traces. In *Under submission*, 2020.

[2] Fatjon Zogaj, José P. Cambronero, Martin Rinard, and Jürgen Cito. Doing More with Less: Characterizing Dataset Downsampling for AutoML. In *Under submission*, 2020.

[3] José P. Cambronero, Raul Castro Fernandez, and Martin Rinard. wranglesearch: Mining Data Wrangling Functions from Python Programs. In *Under submission*, 2021.

[4] Limor Appelbaum, José P. Cambronero, and et al. Development and Validation of a Pancreatic Cancer Risk Model for the General Population Using Electronic Health Records: An Observational Study. In *European Journal of Cancer*, 2020.

[5] José P. Cambronero, Jürgen Cito, and Martin Rinard. AMS: Generating AutoML Search Spaces from Weak Specifications. In *ESEC/FSE*, 2020.

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

[7] 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.

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

[9] 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.

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

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 OOPSLA 2020**
- **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 - Spring 2020)**: I served as Treasurer on the CSAIL Student Committee. I managed the group's budget and contributed to the organization of social events, such as a weekly event featuring baked goods and socializing among graduate students in CSAIL.

Mentoring/Advising

- Jennifer McCleary (MIT) MEng Thesis: pancreatic cancer risk modeling (Fall 2019 - January 2020)
- Alex Berg (MIT) Undergraduate research: pancreatic cancer risk modeling (Summer 2020)
- Thomas Xiong (MIT) MEng Thesis: pancreatic cancer risk modeling (Fall 2020 - Spring 2021)
- Lori Zhang (MIT) Undergraduate research: pancreatic cancer risk modeling (Summer 2020 - Spring 2021)

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