RESEARCH STATEMENT

Marco Aurélio Gerosa
Department of Computer Science
University of São Paulo (USP)
R. do Matão 1010, São Paulo, Brazil
gerosa@ime.usp.br

My research interests lie at the intersection of data science, software engineering, and social computing. I enthusiastically analyze open source communities and their developers using quantitative data mining and statistical techniques as well as qualitative interviews, surveys, and archival analysis, and then develop tools to enhance those communities. Consider the recent work conducted by my former Ph.D. student, Igor Steinmacher, who graduated this year. We quantitatively analyzed software repositories to identify factors that lead newcomer developers to dropout, and then we conducted systematic reviews, surveys, and interviews to model the barriers that newcomers face. The results were published at the *18th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW 2015)*, the field's main conference. Based on the conceived model, we developed a portal (http://www.flosscoach.com) that reduces the identified barriers by providing awareness information for newcomers. Our paper describing the portal analysis was accepted for publication at ICSE 2016 (*International Conference on Software Engineering*). My postdoc Christoph Treude (http://ctreude.ca) is currently applying natural language processing techniques to the large amount of available software repository and social media data to feed the portal. We are currently integrating the portal into the Federal Government public software portal, which will facilitate new developer contribution to open source software projects used by the Brazilian government. Besides the impact on practice, this line of research generated several scientific publications, as my CV attests.

In other work, I used a big data/data science approach to analyze source code and version control systems. My current Ph.D. student Igor Wiese mines social data to build prediction models that recommend software artifacts that should be changed together. Partial results have been published in *IEEE Latin American Transactions* and several area conferences. Another Ph.D. student of mine, Gustavo Oliva, works on a related theme, mining several aspects of the source code structure to correlate to co-changes. We co-authored a chapter in the recently published book *The Art and Science of Analyzing Software Data* (http://www.amazon.com/The-Science-Analyzing-Software-Data/dp/0124115195), explaining the concepts, techniques, and tools used to identify change coupling. Ultimately, we aim to develop a software analytics tool for an Integrated Development Environment that supports engineers in their work. As a base for this tool and our other studies that mine software repositories, we are developing a tool named MetricMiner (http://www.metricminer.org.br), which supports software data metrics data collection and processing.

Besides my work related to the analysis of software engineering data, I also conducted work related to education data and social media analysis (e.g. Twitter), as can be seen in my CV. My education data analysis work was done in collaboration with my Ph.D. advisor's group and was published in the *Computers and Education* Journal. We analyzed educational forum data to provide alerts to course mediators in order to provide timely feedback. In collaboration with a data security specialist colleague from the department, the work on Twitter evaluated the use of Twitter messages as a source for security alerts. This collaboration began when the colleague's students attended my course Special Topics on Web Development, in which I covered the techniques of collecting, wrangling, and analyzing social media data on the Web. In that class, half of the students ended up publishing their term project paper in a peer-reviewed venue.

I also coordinate software engineering projects aimed at developing social systems on the Web. Along with colleagues from the Architecture School (FAU) and the Communication and Arts School (ECA) at the University of São Paulo, we received grant funds to establish the *Collaborative Environments on the Web Research Center.* This center, which I coordinated from 2013 to 2015, comprises 12 faculty members, 23 graduate students, and 19 undergraduate students from the three schools. We are developing two projects that received several awards: a social networking site for sharing and analyzing architecture images (http://www.arquigrafia.org.br, in Portuguese), and a mobile app for improving the mobility of visually impaired people (Smart Audio City Guide). We are also starting a project on Smart Cities, funded by HP Brazil, in which I am the principal investigator. Aiming to use a big data approach to support smart cities, we are analyzing smart cities scenarios and developing simulators, as well as engaging in the HP effort on real-time event processing systems on top of Apache Storm and Spark, which are frequently used for big data analytics.

I adopt the following approaches as a researcher:

- Pragmatism – I use whatever method/approach I think most suitable for the problem and context at hand. I have used mixed quantitative and qualitative techniques through numerous methods, such as case study, experiment, action research, grounded theory, systematic review, etc.
- Engineering – I like to view real-world problems from an engineering perspective, and develop tools and applications for them. I spend a lot of time reflecting and designing solutions.
- Innovation – I love to do things differently and thrive on new ideas.
- Quality – I pursue quality in everything I do.
- Multiple-areas – I have many interests and, moreover, a passion for exploring new areas and the consequent dissonance that reveals many problems and opportunities and creates a flow of ideas. I like to bridge and interconnect different knowledge domains.

In the future, besides expanding the current research lines, I plan to explore new ways of analyzing sociotechnical data to improve software engineering, and to facilitate end-user programming and new developers' project entrance.