Fernando Chirigati

Curriculum Vitæ

Contact Information

NYU Tandon School of Engineering ∘ CSE Department
2 MetroTech Center, 10th Floor ∘ Brooklyn, NY 11201, USA
✉ fchirigati@nyu.edu
🌐 http://fchirigati.com

Education

Jan 2012 – Present | **Ph.D., Computer Science**, *NYU Tandon School of Engineering, Brooklyn – USA*.
Advisor: Juliana Freire, Ph.D.

Jan 2007 – Dec 2011 | **B.E., Computer and Information Engineering**, *Federal University of Rio de Janeiro, Rio de Janeiro – Brazil*.
Advisor: Marta Mattoso, D.Sc.

Aug 2010 – Dec 2010 | **One-Semester Exchange Program**, *University of Central Florida, Orlando – USA*.
Scholarship granted by the Brazilian government.

Research Interests

Data Management ∘ Data Mining ∘ Large-Scale Data Analytics
Provenance Management and Analytics ∘ Reproducibility ∘ Data Visualization

Research Experience

Jan 2012 – Present | **Research Assistant**
NYU Tandon School of Engineering, Brooklyn – USA
Supervisor: Juliana Freire

May 2016 – Aug 2016 | **Summer Research Intern**
New York Structured Data Research Group
Google Research, New York City – USA
Supervisors: Flip Korn and Cong Yu

May 2015 – Aug 2015 | **Summer Research Intern**
New York Structured Data Research Group
Google Research, New York City – USA
Supervisors: Flip Korn and Cong Yu

Jun 2012 – Jan 2015 | **DataONE Working Group Member**
Scientific Workflows and Provenance Working Group

Jun 2013 – Aug 2013 | **Summer Research Intern**
IBM T. J. Watson Research Center, Yorktown Heights – USA
Supervisors: Jérôme Siméon and Martin Hirzel

Jan 2009 – Feb 2009	**Research Intern** University of Utah, Salt Lake City – USA Supervisors: Juliana Freire and Cláudio T. Silva
Aug 2007 – Aug 2010	**Research Assistant** Federal University of Rio de Janeiro, Rio de Janeiro – Brazil Supervisor: Marta Mattoso

RESEARCH PROJECTS

Aug 2014 – Present	**Urban Data Management and Analytics** NYU Tandon School of Engineering, Brooklyn – USA *Description*: Discovering relationships between spatio-temporal urban datasets can lead to many insights and help understand how the city behaves. Given the sheer number and size of the data sets, and the diverse spatial and temporal scales at which the data is available, this task presents computational challenges on all fronts, from indexing and querying to analyzing the relationships. Also, it is non-trivial to differentiate between meaningful and spurious relationships. This project takes a first step towards addressing these challenges and proposing a solution that is scalable and effective at identifying potentially meaningful relationships.
Jan 2012 – Present	**Reproducibility in Science** NYU Tandon School of Engineering, Brooklyn – USA *Description*: While reproducibility is a core component of the scientific process, science falls far short of reproducible results. Most computational experiments are specified only informally in papers, where experimental results are briefly described in figure captions; the code that produced the results is seldom available; and configuration parameters change results in unforeseen ways. This project is focused on developing a suite of tools and infrastructure that supports the process of sharing, testing, and re-using scientific experiments and results. In particular, the goal is to assist researchers in streamlining their research process to make their data interoperable and reproducible. ReproZip and noWorkflow are examples of outcomes of the project.
Jul 2009 – Aug 2010	**Management of Large-Scale Scientific Experiments** Federal University of Rio de Janeiro, Rio de Janeiro – Brazil *Description*: Many different scientific areas, including deep-sea oil exploitation and bioinformatics, require simulating large-scale experiments. These simulations must be designed efficiently to avoid the overconsumption of computational resources, as this can become incredibly expensive. In this project, the focus is on building tools to design scientific workflows for such simulations, from its conception to its execution and provenance analysis. The use of provenance data is key in this project: by analyzing ongoing and past executions, previous results can be re-used, thus avoiding expensive, redundant computations.
Aug 2007 – Jul 2009	**Data Provenance in Scientific Experiments** Federal University of Rio de Janeiro, Rio de Janeiro – Brazil *Description*: As the amount of data becomes massive and the nature of the scientific workflows becomes more heterogeneous, with data and tasks located in a distributed manner, the need for integrating provenance from the different experiment steps becomes more evident. The goal of this project is to propose efficient solutions, developed in scientific workflow systems, to store and analyze such provenance data, while integrating the different pieces into a single queryable store.

Awards and Honors

Honorable Mention **Best Demonstration – SIGMOD 2017**
Querying and Exploring Polygamous Relationships in Urban Spatio-Temporal Data Sets
2017

Award **Most Reproducible Paper – SIGMOD 2017**
Data Polygamy: The Many-Many Relationships among Urban Spatio-Temporal Data Sets
2017

Award **Student Travel Award – SIGMOD 2017**
2017

Award **Pearl Brownstein Doctoral Research Award – NYU Tandon**
Doctoral research that shows the greatest promise.
2016

2ⁿᵈ Place **Programming Contest – SIGMOD 2014**
Together with Tuan-Anh Hoang-Vu, Kien Pham, and Huy T. Vo.
2014

Award **Deborah Rosenthal, MD Award – NYU Tandon**
Outstanding performance on the Ph.D. qualifying examination.
2014

Honorable Mention **A3P Special Honor – Federal University of Rio de Janeiro**
Outstanding performance achieved at the Federal University of Rio de Janeiro, given by the Alumni Association of the Polytechnic School (A3P).
2013

Honorable Mention **Magna Cum Laude Honor – Federal University of Rio de Janeiro**
Outstanding performance achieved in Computer and Information Engineering.
2013

Honorable Mention **Research Honor – Federal University of Rio de Janeiro**
Given by the Academic Deliberative Council of Graduate Department of Engineering (COPPE).
2010

Award **Best Poster – XXIV Brazilian Symposium on Databases**
Development of Explicit Control Structures for SWfMS VisTrails (in Portuguese)
2009

Honorable Mention **Best Presentation – Federal University of Rio de Janeiro**
Top 10 presentations among more than 500 presentations, during the XXXI Conference on Young Research Assistant.
2010

Publications

Journals

2016   *ReproZip: The Reproducibility Packer*, R. Rampin, **F. Chirigati**, D. Shasha, J. Freire, and V. Steeves. In Journal of Open Source Software (**JOSS**), 2016

Knowledge Exploration Using Tables on the Web, **F. Chirigati**, J. Liu, F. Korn, Y. Wu, C. Yu, and H. Zhang. In Proceedings of the VLDB Endowment (**PVLDB**), 10(3), pp. 193-204, 2016

Exploring What not to Clean in Urban Data: A Study Using New York City Taxi Trips, J. Freire, A. Bessa, **F. Chirigati**, H. T. Vo, and K. Zhao. In IEEE Data Engineering Bulletin, 39(2), pp. 63-77, 2016

2015   *YesWorkflow: A User-Oriented, Language-Independent Tool for Recovering Workflow Information from Scripts*, T. McPhillips, T. Song, T. Kolisnik, S. Aulenbach, K. Belhajjame, R. Kyle Bocinsky, Y. Cao, J. Cheney, **F. Chirigati**, S. Dey, J. Freire, C. Jones, J. Hanken, K. W. Kintigh, T. A. Kohler, D. Koop, J. A. Macklin, P. Missier, M. Schildhauer, C. Schwalm, Y. Wei, M. Bieda, B. Ludäscher. In International Journal of Digital Curation (**IJDC**), 10(1), pp. 298-313, 2015

2014   *The More the Merrier: Efficient Multi-Source Graph Traversal*, M. Then, M. Kaufmann, **F. Chirigati**, T. Hoang-Vu, K. Pham, A. Kemper, T. Neumann, and H. T. Vo. In Proceedings of the VLDB Endowment (**PVLDB**), 8(4), pp. 449-460, 2014

The PBase Scientific Workflow Provenance Repository, V. Cuevas-Vicenttín, P. Kianmajd, B. Ludäscher, P. Missier, **F. Chirigati**, Y. Wei, D. Koop, and S. Dey. In International Journal of Digital Curation (**IJDC**), 9(2), pp. 28-38, 2014

2013   *A Computational Reproducibility Benchmark*, **F. Chirigati**, M. Troyer, D. Shasha, and J. Freire. In IEEE Data Engineering Bulletin, 36(4), pp. 54-59, 2013

Chiron: A Parallel Engine for Algebraic Scientific Workflows, E. Ogasawara, J. Dias, V. Souza, **F. Chirigati**, D. Oliveira, F. Porto, P. Valduriez, and M. Mattoso. In Journal of Concurrency and Computation: Practice and Experience, 25(16), pp. 2327-2341, 2013

2011   *Similarity-Based Workflow Clustering*, V. Souza, **F. Chirigati**, K. Maia, E. Ogasawara, D. Oliveira, V. Braganholo, L. Murta, and M. Mattoso. In Journal of Computational Interdisciplinary Sciences, vol. 2, pp. 23-35, 2011

Conferences and Workshops

2017   *Querying and Exploring Polygamous Relationships in Urban Spatio-Temporal Data Sets*, Y. Chan, **F. Chirigati**, H. Doraiswamy, C. Silva and J. Freire. In Proceedings of the 2017 ACM SIGMOD International Conference on Management of Data (**SIGMOD**), pp. 1643-1646, 2017 **Honorable Mention, SIGMOD Best Demonstration Award**

2016   *Data Polygamy: The Many-Many Relationships among Urban Spatio-Temporal Data Sets*, **F. Chirigati**, H. Doraiswamy, T. Damoulas, and J. Freire. In Proceedings of the 2016 ACM SIGMOD International Conference on Management of Data (**SIGMOD**), pp. 1011-1025, 2016 **SIGMOD Most Reproducible Paper Award**

ReproZip: Computational Reproducibility With Ease, **F. Chirigati**, R. Rampin, D. Shasha, and J. Freire. In Proceedings of the 2016 ACM SIGMOD International Conference on Management of Data (**SIGMOD**), pp. 2085-2088, 2016

Virtual Lightweight Snapshots for Consistent Analytics in NoSQL Stores, **F. Chirigati**, J. Siméon, M. Hirzel, and J. Freire. In Proceedings of the 32nd International Conference on Data Engineering (**ICDE**), pp. 1310-1321, 2016

2015   *noWorkflow: Capturing and Analyzing Provenance of Scripts*, L. Murta, V. Braganholo, **F. Chirigati**, D. Koop, and J. Freire. In Provenance and Annotation of Data and Processes, vol. 8628, Lecture Notes in Computer Science (**LNCS**), pp. 71-83, Springer International Publishing, 2015

2013   *Packing Experiments for Sharing and Publication*, **F. Chirigati**, D. Shasha, and J. Freire. In Proceedings of the 2013 International Conference on Management of Data (**SIGMOD**), pp. 977-980, 2013

ReproZip: Using Provenance to Support Computational Reproducibility, **F. Chirigati**, D. Shasha, and J. Freire. In Proceedings of the 5th USENIX Conference on Theory and Practice of Provenance (**TaPP**), 2013

VisTrails Provenance Traces for Benchmarking, **F. Chirigati**, D. Koop, J. Freire, and C. Silva. In Proceedings of the 2013 Joint **EDBT/ICDT** Workshops, pp. 323-324, 2013

2012   *Towards Integrating Workflow and Database Provenance*, **F. Chirigati** and J. Freire. In Provenance and Annotation of Data and Processes, vol. 7525, Lecture Notes in Computer Science (**LNCS**), pp. 11-23, Springer Berlin / Heidelberg, 2012

Evaluating Parameter Sweep Workflows in High Performance Computing, **F. Chirigati**, V. Souza, E. Ogasawara, D. Oliveira, J. Dias, F. Porto, P. Valduriez, and M. Mattoso. In Proceedings of the 1st International Workshop on Scalable Workflow Enactment Engines and Technologies (**SWEET**), article 2, 2012

2011   *An Evaluation of the Distribution of Dynamic and Static Activities in Parallel Environments using Hydra*, V. Souza, **F. Chirigati**, E. Ogasawara, J. Dias, D. Oliveira, F. Porto, P. Valduriez, and M. Mattoso. In Proceedings of the XXXI Congress of the Brazilian Computer Society, 2011
In Portuguese

2010   *SimiFlow: An Architecture for Clustering Workflows by Similarity*, V. Souza, **F. Chirigati**, K. Maia, E. Ogasawara, D. Oliveira, V. Braganholo, L. Murta, and M. Mattoso. In Proceedings of the XXX Congress of the Brazilian Computer Society, 2010
In Portuguese

GExpLine: A Tool for Supporting Experiment Composition, D. Oliveira, E. Ogasawara, **F. Chirigati**, V. Souza, L. Murta, and M. Mattoso. In Provenance and Annotation of Data and Processes, vol. 6378, Lecture Notes in Computer Science (LNCS), pp. 251-259, Springer Berlin / Heidelberg, 2010

2009   *A Semantic Approach for Scientific Experiment Lines using Ontologies*, D. Oliveira, E. Ogasawara, **F. Chirigati**, V. Souza, L. Murta, C. Werner, and M. Mattoso. In Proceedings of the III e-Science Workshop, XXIV Brazilian Symposium on Databases, 2009
In Portuguese

Scientific Workflow Management System Applied to Uncertainty Quantification in Large Eddy Simulation, G. Guerra, F. Rochinha, R. Elias, A. Coutinho, V. Braganholo, D. Oliveira, E. Ogasawara, **F. Chirigati**, and M. Mattoso. In Proceedings of the 30th Iberian-Latin-American Congress on Computational Methods in Engineering (CILAMCE), 2009

Exploring Many Task Computing in Scientific Workflows, E. Ogasawara, D. Oliveira, **F. Chirigati**, C. E. Barbosa, R. Elias, V. Braganholo, A. Coutinho, and M. Mattoso. In Proceedings of the 2nd Workshop on Many-Task Computing on Grids and Supercomputers, International Conference for High Performance, Networking, Storage and Analysis (SC), 2009

2008   *Expliciting Control Flow in Scientific Workflows*, S. M. S. Cruz, **F. Chirigati**, R. Dahis, M. L. M. Campos, and M. Mattoso. In Proceeding of the II e-Science Workshop, XXIII Brazilian Symposium on Databases, 2008

Using Explicit Control Processes in Distributed Workflows to Gather Provenance, S. M. S. Cruz, **F. Chirigati**, R. Dahis, M. L. M. Campos, and M. Mattoso. In Provenance and Annotation of Data and Processes, vol. 5272, Lecture Notes in Computer Science (**LNCS**), pp. 186-199, Springer Berlin / Heidelberg, 2008

Book Chapters

2017   *Glossary*, A. Rokem and **F. Chirigati**. In J. Kitzes, D. Turek, and F. Deniz (Eds.), The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences, 2017

Provenance and Reproducibility, **F. Chirigati** and J. Freire. In L. Liu and M. T. Özsu (Eds.), Encyclopedia of Database Systems, 2017

2014   *Reproducibility Using VisTrails*, J. Freire, D. Koop, **F. Chirigati**, and C. Silva. In V. Stodden, F. Leisch, and R. Peng (Eds.), Implementing Reproducible Computational Research (The R Series), 2014

Editorials

2016   *A Collaborative Approach to Computational Reproducibility*, **F. Chirigati**, R. Capone, R. Rampin, J. Freire, and D. Shasha. In Information Systems, vol. 59, pp. 95-97, 2016

Posters

2016   *Enhancing Scholarly Communication with ReproZip*, **F. Chirigati**, R. Rampin, V. Steeves, D. Shasha, and J. Freire. FORCE2016 Conference, 2016

2014   *Constructing a Social Network Analysis System for SIGMOD 2014 Programming Contest*, **F. Chirigati**, K. Pham, T. Hoang-Vu, and H. T. Vo. SIGMOD 2014 Programming Contest, 2014

Provenance Storage, Querying, and Visualization in PBase, V. Cuevas-Vicenttín, P. Kianmajd, B. Ludäscher, P. Missier, **F. Chirigati**, Y. Wei, D. Koop, and S. Dey. In Proceedings of the International Provenance and Annotation Workshop (**IPAW**), 2014

2013   *ReproZip: Packing Experiments for Sharing and Publication*, **F. Chirigati**, D. Shasha, and J. Freire. Beyond the PDF 2 Conference, 2013

2009   *Procedure to Build Scientific Workflows*, M. P. Rodrigues, J. C. C. Fernandez, **F. Chirigati**, S. M. S. Cruz, and M. C. R. Cavalcanti. XXIV Brazilian Symposium on Databases, 2009
In Portuguese

Development of Explicit Control Structures for SWfMS VisTrails, **F. Chirigati**, R. Dahis, S. M. S. Cruz, J. Freire, C. Silva, and M. Mattoso. XXIV Brazilian Symposium on Databases, 2009
Best Poster Award
In Portuguese

A Conception Process for Abstract Workflows: An Example on Deep Water Oil Exploitation Domain, W. Martinho, E. Ogasawara, D. Oliveira, **F. Chirigati**, F. Correa, B. Jacob, I. Santos, G. H. Travassos, and M. Mattoso. 5th IEEE International Conference on e-Science, 2009

Reproducibility Papers

2017   *HESML: A Scalable Ontology-based Semantic Similarity Measures Library with a Set of Reproducible Experiments and a Replication Dataset*, J. Lastra-Díaz, A. García-Serrano, M. Batet, M. Fernández, and **F. Chirigati**. In Information Systems, vol. 66, pp. 97-118, 2017

2016   *Reproducible Experiments on Dynamic Resource Allocation in Cloud Data Centers*, A. Wolke, M. Bichler, **F. Chirigati**, and V. Steeves. In Information Systems, vol. 59, pp. 98-101, 2016

Presentations

2017   **Data Polygamy: The Many-Many Relationships among Urban Spatio-Temporal Data Sets**
Invited Talk at University of Münster
Münster, Germany, 2017

Preserving and Reproducing Research with ReproZip
Invited Talk at Brainhack NYC
New York City, USA, 2017

2016   **Data Polygamy: The Many-Many Relationships among Urban Spatio-Temporal Data Sets**
AWS re:Invent 2016
Las Vegas, USA, 2016

Preserving and Reproducing Research with ReproZip
Preservation and Archiving Special Interest Group (PASIG), Fall 2016 Meeting
New York City, USA, 2016

Data Polygamy: The Many-Many Relationships among Urban Spatio-Temporal Data Sets
International Conference on Management of Data (SIGMOD)
San Francisco, USA, 2016

Virtual Lightweight Snapshots for Consistent Analytics in NoSQL Stores
32nd International Conference on Data Engineering (ICDE)
Helsinki, Finland, 2016

ReproZip: Computational Reproducibility with Ease
Dagstuhl Seminar 16041, Reproducibility of Data-Oriented Experiments in e-Science
Wadern, Germany, 2016

2015   **Achieving Reproducibility with ReproZip**
Invited talk at the CS Colloquium, Columbia University
New York City, USA, 2015

Facilitating Reproducibility After the Fact
BIDS Reproducibility Conference, University of California, Berkeley
Berkeley, USA, 2015

Facilitating Reproducibility After the Fact
Reproducibility Seminar, eScience Institute, University of Washington
Seattle, USA, 2015

2014   **Constructing a Social Network Analysis System for SIGMOD 2014 Programming Contest**
SIGMOD Programming Contest
Snowbird, USA, 2014

2013   **ReproZip: Packing Experiments for Sharing and Publication**
Workshop on Software Infrastructure for Reproducibility in Science
Brooklyn, USA, 2013

ReproZip: Packing Experiments for Sharing and Publication
Beyond the PDF 2 Conference – Visions of the Future Session
Amsterdam, Netherlands, 2013

2012   **ReproZip: Packing Experiments for Sharing and Publication**
ICERM Workshop on Reproducibility in Computational and Experimental Mathematics
Providence, USA, 2012

Towards Integrating Workflow and Database Provenance
4th International Provenance and Annotation Workshop (IPAW)
Santa Barbara, USA, 2012

Evaluating Parameter Sweep Workflows in High Performance Computing
1st International Workshop on Scalable Workflow Enactment Engines and Technologies (SWEET)
Scottsdale, USA, 2012

Professional Activities

2015 – Present   **Reproducibility Editor**
Information Systems Journal, Elsevier North-Holland

2018   **Program Committee Member – Research Track**
Very Large Data Bases (VLDB) Conference, 2018

Program Committee Member – Demo Track
ACM International Conference on Management of Data (SIGMOD), 2018

2017	**Program Committee Member** SciPy Conference, 2017
2016	**Reproducibility Committee Member** ACM International Conference on Management of Data (SIGMOD), 2016
	Artifact Evaluation Committee Member European Conference on Object-Oriented Programming (ECOOP), 2016
2015	**Reproducibility Committee Member** ACM International Conference on Management of Data (SIGMOD), 2015
	Artifact Evaluation Committee Member European Conference on Object-Oriented Programming (ECOOP), 2015
2013	**Student Volunteer** ACM International Conference on Management of Data (SIGMOD), 2013
	Co-Organizer Workshop on Software Infrastructure for Reproducibility in Science, Brooklyn, USA, 2013
	Co-Organizer Workshop on Reproducibility in Science, Brooklyn, USA, 2013

Freely-Available Software Systems

ReproZip

https://www.reprozip.org/

ReproZip is a tool that automatically captures provenance of experiments and packs all the necessary files, library dependencies, and variables to reproduce the results. Reviewers can then unpack and run the experiments without having to install any additional software.

noWorkflow

http://gems-uff.github.io/noworkflow/

noWorkflow is a tool that can transparently capture detailed provenance information from Python scripts. It is non-intrusive, does not require users to change the way they work, and provides different ways to analyze the captured provenance.