Gianmarco De Francisci Morales

Curriculum Vitae *(July 28, 2014)*

Experience

Sep '13–Present **Research Scientist**, *Yahoo Labs*, Barcelona, Spain.
Research on web mining and data intensive scalable computing.

Aug '11–Present **Apache Pig Committer**, *Apache Software Foundation*.
Committer for the Apache Pig project from the Hadoop ecosystem. Member of the open source community that develops Pig, a framework for large scale data analysis.

Oct '11–Sep '13 **Postdoctoral Researcher**, *Yahoo! Research*, Barcelona, Spain.
Research on web and social network mining.

Oct '12–Dec '12 **Visiting Scientist**, *Yahoo! IntoNow*, Sunnyvale, USA.
Design and development of a backend for stream processing based on Storm.

Apr '12–Aug '12 **Google Summer of Code Mentor**, *Google & Apache Software Foundation*.
Mentor for Apache Pig in the GSoC initiative. Implementation of a distributed RANK function in MapReduce. Integrated in Pig 0.11. code

Nov '09–Jun '12 **Research Associate**, *National Research Council (ISTI-CNR)*, Pisa, Italy.
Research on cloud computing for large scale data analysis.

Oct '10–Sep '11 **Research Intern**, *Yahoo! Research*, Barcelona, Spain.
Research on data intensive scalable computing.

May '11–Aug '11 **Google Summer of Code Student**, *Google & Apache Software Foundation*.
Student in the GSoC project for Apache Pig. Sugar for Pig: syntactic sugar features to simplify the use of Pig (Java, ANTLR). Integrated in Pig 0.10. code 1 code 2

May '10–Aug '10 **Google Summer of Code Student**, *Google & Apache Software Foundation*.
Student in the GSoC project for Apache Pig. Design and implementation of a binary comparator for secondary sort. The new comparator is 10x faster, and produces a 20% end-to-end improvement for relevant queries (Java). Integrated in Pig 0.9. code

May '07–Feb'09 **System Administrator**, *Consorzio COMETA*, Catania, Italy.
Administration of Linux clusters for High Performance Computing (HPC). Management of grid middleware (gLite). Porting, integration and development of grid-enabled and parallel (MPI) applications.

Apr '04–Jul '04 **Intern**, *National Institute for Nuclear Physics (INFN)*, Catania, Italy.
Design and development of a framework for booking and allocation of resources in a Grid environment (Java).

(+34) 672 199131 • gdfm@gdfm.me • gdfm.me

Education

Mar '09–Mar'12 **Ph.D. ("Doctor Europaeus") in Computer Science and Engineering**, *IMT Institute for Advanced Studies Lucca*, Italy, Full marks and honors (*Excellent*).
Thesis: "Big Data and the Web: Algorithms for Data Intensive Scalable Computing".

Oct '04–Apr '08 **Master's in Computer Engineering**, *University of Catania*, Italy, Full marks and honors (*110/110 cum laude*).
Thesis: "Multi-objective design space exploration of embedded systems in a Grid environment" (developed in C++, written in English).

Mar '06–Aug '06 **Erasmus student**, *Pazmany Peter Catholic University*, Budapest, Hungary.
Artificial Intelligence. Design Patterns in Java. Cellular Neural Networks.

Oct '01–Oct '04 **Bachelor's in Computer Engineering**, *University of Catania*, Italy, Full marks and honors (*110/110 cum laude*).
Thesis: "Serivzi di allocazione e monitoraggio dei job e delle risorse di una Grid" (Allocation and monitoring services for jobs and resources in a Grid environment, Java).

Courses

Oct '11–Dec '11 **Introduction to Artificial Intelligence**, *Stanford University*, ai-class.com.
First Massive Online Open Course (MOOC) on Artificial Intelligence, by Sebastian Thrun and Peter Norvig.

Oct '11–Dec '11 **Introduction to Machine Learning**, *Stanford Univesrity*, ml-class.com.
First Massive Online Open Course (MOOC) on Machine Learning, by Andrew Ng.

Jul '10 **Lipari Summer School on Computational Complex Systems**, *Jacob T. Schwartz International School for Scientific Research*, Lipari, Italy.
Social Network Mining: Graph mining; Social network analysis; Complex networks.

Jul '09 **CASE Summer School**, *Free University of Bozen*, Bolzano, Italy.
Software Engineering: Agile software process management; Test Driven Development (TDD); Continuous testing of configurable systems; Software and process measurement.

Dec '08 **Matlab Official Course**, *The MathWorks*, Palermo, Italy.
Matlab: Fundamentals; GUI design; Integration with external languages; Distributed and parallel computing. (ML01 - ML04 - ML05 - DC01).

Awards

2010 Erasmus student mobility scholarship.
2009 First in admission ranking for Ph.D. out of 446 applicants.
2009 Ph.D. scholarship for 3 years sponsored by the European Union.
2008 Award for outstanding academic career from University of Catania.
2006 Erasmus scholarship from ERSU (Regional Agency for University Studies).
2006 Socrates mobility scholarship from Univeristy of Catania.
2005 Master's scholarship from ERSU (Regional Agency for University Studies).
2004 Award for achieving full marks and honors in minimum time from ERSU.
2003 Bachelor's scholarship from ERSU (Regional Agency for University Studies).
2002 Bachelor's scholarship from ERSU (Regional Agency for University Studies).

(+34) 672 199131 • gdfm@gdfm.me • gdfm.me

Languages

Italian	**Native**	
English	**Proficient**	*European level C2*
Spanish	**Proficient**	*European level C2*
Japanese	**Basic**	*European level A2*

Computer skills

Languages	Java, C/C++, Python, Perl, Bash, SQL, LaTeX.
Big Data	Hadoop ecosystem, Pig, Giraph, Grafos, S4, Storm.
Tools	Matlab, ANTLR, Subversion, git, ant+ivy, Maven.

Interests

- Chess, reading, cooking, archery, basketball, photography.

Patents

- Two methods for matching social content in map-reduce architecture.
- An automatic method to build flowcharts for e-shopping recommendations.
- An information retrieval method for online news suggestion using closed captions.
- A system for recommending relevant Web content to second-screen application users.

Invited Talks

"SAMOA: Scalable Advanced Massive Online Analysis", NoSQL Matters '13, Barcelona, 30 Nov 2013.

"SAMOA: A Platform for Mining Big Data Streams," RAMSS '13 Keynote, Rio De Janeiro, 14 May 2013.

"Big Data and the Web," IMT Institute for Advanced Studies, Lucca, 21 March 2012.

"How to Survive the Data Deluge: Petabyte Scale Cloud Computing," ISTI-CNR, Pisa, 18 January 2010.

Publications

R. Blanco, G. De Francisci Morales, and F. Silvestri, "IntoNews: Online News Retrieval using Closed Captions," *IP&M: Information Processing & Management*, 2014.

B. Thomee and G. De Francisci Morales, "Automatic Discovery of Global and Local Equivalence Relationships in Labeled Geo-Spatial Data," in *HT '14: 25th ACM Conference on Hypertext and Social Media*, 2014.

D. Marron, A. Bifet, and G. De Francisci Morales, "Random Forests of Very

Fast Decision Trees on GPU for Mining Evolving Big Data Streams," in *ECAI '14: 21st European Conference on Artificial Intellingence*, 2014.

C. Castillo, G. De Francisci Morales, M. Mendoza, and N. Khan, "Says who? Automatic Text-Based Content Analysis of Television News," in *MNLP '13: 1st Workshop on Mining unstructured big data using Natural Language Processing @CIKM*, pp. 53–60, 2013.

M. Das, G. De Francisci Morales, A. Gionis, and I. Weber, "Learning to Question: Leveraging User Preferences for Shopping Advice," in *KDD '13: 19th ACM Conference on Knowledge Discovery and Data Mining*, pp. 203–211, 2013.

F. Bonchi, G. De Francisci Morales, A. Gionis, and A. Ukkonen, "Activity Preserving Graph Simplification," *DMKD: Data Mining and Knowledge Discovery*, vol. 27, no. 3, pp. 321–343, 2013.

C. Castillo, G. De Francisci Morales, and A. Shekhawat, "Online Matching of Web Content to Closed Captions in IntoNow," in *SIGIR '13: 36th International Conference on Research and Development in Information Retrieval*, pp. 1115–1116, (demo) 2013.

G. De Francisci Morales, "SAMOA: A Platform for Mining Big Data Streams," in *RAMSS'13: 2nd International Workshop on Real-Time Analysis and Mining of Social Streams @WWW*, pp. 777–778, (extended abstract) 2013.

R. Blanco, G. De Francisci Morales, and F. Silvestri, "Towards Leveraging Closed Captions for News Retrieval," in *WWW '13: 22nd International Conference on World Wide Web*, pp. 135–136, (poster) 2013.

G. De Francisci Morales and A. Shekhawat, "The Future of Second Screen Experience," in *Workshop on Exploring and Enhancing the User Experience for Television @CHI*, 2013.

I. Bordino, G. De Francisci Morales, I. Weber, and F. Bonchi, "From Machu_Picchu to "rafting the urubamba river": Anticipating information needs via the Entity-Query Graph," in *WSDM '13: 6th International Conference on Web Search and Data Mining*, pp. 275–284, 2013.

G. De Francisci Morales, A. Gionis, and C. Lucchese, "From Chatter to Headlines: Harnessing the Real-Time Web for Personalized News Recommendation," in *WSDM '12: 5th International Conference on Web Search and Data Mining*, pp. 153–162, 2012.

G. De Francisci Morales, A. Gionis, and M. Sozio, "Social Content Matching in MapReduce," *VLDB Endowment*, vol. 4, no. 7, pp. 460–469, 2011.

R. Baraglia, G. De Francisci Morales, and C. Lucchese, "Document Similarity Self-Join with MapReduce," in *ICDM '10: 10th International Conference on Data Mining*, pp. 731–736, 2010.

G. De Francisci Morales, C. Lucchese, and R. Baraglia, "Scaling Out All Pairs Similarity Search with MapReduce," in *LSDS-IR '10: 8th Workshop on Large-Scale Distributed Systems for Information Retrieval @SIGIR*, pp. 25–30, 2010.

R. Baraglia, C. Lucchese, and G. De Francisci Morales, "Large-scale Data Analysis on the Cloud," in *XXIV Convegno Annuale del CMG-Italia*, (best paper award) 2010.

G. De Francisci Morales, "Cloud Computing for Large Scale Data Analysis," tech. rep., IMT Institute for Advanced Studies, February 2010.

C. M. Iacono-Manno, M. Fargetta, R. Barbera, A. Falzone, G. Andronico, S. Monforte, A. Muoio, R. Bruno, P. D. Primo, S. Orlando, E. Leggio, A. Lombardo, G. Passaro, G. De Francisci Morales, and S. Blandino, "The Sicilian Grid Infrastructure for High Performance Computing," *International Journal of Distributed Systems and Technologies*, vol. 1, no. 1, pp. 40–54, 2010.

V. Catania, A. G. D. Nuovo, M. Palesi, D. Patti, and G. De Francisci Morales, "An Effective Methodology to Multi-objective Design of Application Domain-specific Embedded Architectures," in *DSD '09: 12th Conference on Digital System Design, Architectures, Methods and Tools*, pp. 643–650, 2009.

V. Catania, G. De Francisci Morales, A. G. D. Nuovo, M. Palesi, and D. Patti, "High Performance Computing for Embedded System Design: A Case Study," in *DSD '08: 11th Conference on Digital System Design Architectures, Methods and Tools*, pp. 656–659, 2008.