Gianmarco De Francisci Morales

Curriculum Vitae *(December 17, 2015)*

Experience

Mar '15–Present **Visiting Scientist**, *Aalto University*, Helsinki, Finland.
Data mining group. Research on graph and news mining.

Sep '13–Mar '15 **Research Scientist**, *Yahoo Labs*, Barcelona, Spain.
Web mining group. 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 Labs*, Barcelona, Spain.
Web mining group. 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.
HPC Lab. 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).

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, Scala, 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 MapReduce 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.

Community Service

Tutorial	"Centrality Measures on Big Graphs: Exact, Approximated, and Distributed Algorithms," @WWW '16, 11 Apr 2016. [web]
Tutorial	"Big Data Stream Mining," @IEEE BigData '14, 27 Oct 2014. [web|slides]
Worhshop	"SNOW '16: 3rd Workshop on Social News on the Web," @WWW, 2016. [web]
Worhshop	"SNOW '14: 2nd Workshop on Social News on the Web," @WWW, 2014. [web]
Workshop	"SNOW '13: 1st Workshop on Social News on the Web," @WWW, 2013. [web]

Presentations

"Mining Big Data Streams: Better Algorithms or Faster Systems?" KAIST, Seoul, 20 Apr 2015.

"SAMOA: A Platform for Mining Big Data Streams" Strata '14, Barcelona, 20 Nov 2014.

"Mining Big Data Streams" Workshop on Online Social Networks: Emerging Trends, University of Cyprus, Nicosia, 8 Oct 2014.

"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.

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

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

Publications

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. "Exploring Controversy in Twitter". In *CSCW '16: 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing*, [demo] 2016.

Muhammad Anis Uddin Nasir, Gianmarco De Francisci Morales, Nicolas Kourtellis, and Marco Serafini. "When Two Choices Are not Enough: Balancing at Scale in Distributed Stream Processing". *arXiv:1510.05714*, 2015.

Muhammad Anis Uddin Nasir, Gianmarco De Francisci Morales, David García-Soriano, Nicolas Kourtellis, and Marco Serafini. "Partial Key Grouping: Load-Balanced Partitioning of Distributed Streams". *arXiv:1510.07623*, 2015.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. "Quantifying Controversy in Social Media". In *WSDM '16: 9th ACM International Conference on Web Search and Data Mining*, 2015.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Mauro Sozio. "Scalable Facility Location for Massive Graphs on Pregel-like Systems". In *CIKM '15: 24th ACM International Conference on Information and Knowledge Management*, pp. 273–282, 2015.

Albert Bifet, Gianmarco De Francisci Morales, Jesse Read, Geoff Holmes, and Bernhard Pfahringer. "Efficient Online Evaluation of Big Data Stream Classifiers". In *KDD '15: 21th ACM International Conference on Knowledge Discovery and Data Mining*, pp. 59–68, 2015.

Nicolas Kourtellis, Gianmarco De Francisci Morales, and Francesco Bonchi. "Scalable Online Betweenness Centrality in Evolving Graphs". *TKDE: IEEE Transaction on Knowledge and Data Engineering*, 27(9):2494–2506, 2015.

Muhammad Anis Uddin Nasir, Gianmarco De Francisci Morales, David García-Soriano, Nicolas Kourtellis, and Marco Serafini. "The Power of Both Choices: Practical Load Balancing for Distributed Stream Processing Engines". In *ICDE '15: 31st IEEE International Conference on Data Engineering*, pp. 137–148, 2015.

Gianmarco De Francisci Morales and Albert Bifet. "SAMOA: Scalable Advanced Massive Online Analysis". *JMLR: Journal of Machine Learning Research*, 16(Jan):149–153, 2015.

Roi Blanco, Gianmarco De Francisci Morales, and Fabrizio Silvestri. "IntoNews: Online News Retrieval using Closed Captions". *IP&M: Information Processing & Management*, 51(1):148–162, 2015.

Albert Bifet and Gianmarco De Francisci Morales. "Big Data Stream Learning with SAMOA". In *ICDM '14: 14th IEEE International Conference on Data Mining*, pp. 1199–1202, [demo] 2014.

Anh Thu Vu, Gianmarco De Francisci Morales, Joao Gama, and Albert Bifet. "Distributed Adaptive Model Rules for Mining Big Data Streams". In *BigData '14: 2nd IEEE International Conference on Big Data*, pp. 345–353, 2014.

Bart Thomee and Gianmarco 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*, pp. 158–168, 2014.

Diego Marron, Albert Bifet, and Gianmarco 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*, pp. 615–620, 2014.

Sandra Gonzalez-Bailon, Gianmarco De Francisci Morales, Marcelo Mendoza, Nasir Khan, and Carlos Castillo. "Cable News Coverage and Online News Stories: A Large-Scale Comparison of Media Bias". *SSRN: Social Science Research Network*, 2389525(Feb), 2014.

Carlos Castillo, Gianmarco De Francisci Morales, Marcelo Mendoza, and Nasir 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.

Mahashweta Das, Gianmarco De Francisci Morales, Aristides Gionis, and Ingmar 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.

Francesco Bonchi, Gianmarco De Francisci Morales, Aristides Gionis, and Antti Ukkonen. "Activity Preserving Graph Simplification". *DMKD: Data Mining and Knowledge Discovery*, 27(3):321–343, 2013.

Carlos Castillo, Gianmarco De Francisci Morales, and Ajay Shekhawat. "Online Matching of Web Content to Closed Captions in IntoNow". In *SIGIR '13: 36th ACM International Conference on Research and Development in Information Retrieval*, pp. 1115–1116, [demo] 2013.

Gianmarco 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.

Roi Blanco, Gianmarco De Francisci Morales, and Fabrizio Silvestri. "Towards Leveraging Closed Captions for News Retrieval". In *WWW '13: 22nd International World Wide Web Conference*, pp. 135–136, [poster] 2013.

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

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

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

Gianmarco De Francisci Morales, Aristides Gionis, and Mauro Sozio. "Social Content Matching in MapReduce". *VLDB Endowment*, 4(7):460–469, 2011.

Ranieri Baraglia, Gianmarco De Francisci Morales, and Claudio Lucchese. "Document Similarity Self-Join with MapReduce". In *ICDM '10: 10th IEEE International Conference on Data Mining*, pp. 731–736, 2010.

Gianmarco De Francisci Morales, Claudio Lucchese, and Ranieri 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.

Ranieri Baraglia, Claudio Lucchese, and Gianmarco De Francisci Morales. "Large-scale Data Analysis on the Cloud". In *XXIV Convegno Annuale del CMG-Italia*, [best paper award] 2010.

Gianmarco De Francisci Morales. "Cloud Computing for Large Scale Data Analysis". Technical report, IMT Institute for Advanced Studies, February 2010.

Carmelo Marcello Iacono-Manno, Marco Fargetta, Roberto Barbera, Alberto Falzone, Giuseppe Andronico, Salvatore Monforte, Annamaria Muoio, Riccardo Bruno, Pietro Di Primo, Salvatore Orlando, Emanuele Leggio, Alessandro Lombardo, Gianluca Passaro, Gianmarco De Francisci Morales, and Simona Blandino. "The Sicilian Grid Infrastructure for High Performance Computing". *International Journal of Distributed Systems and Technologies*, 1(1):40–54, 2010.

Vincenzo Catania, Alessandro G. Di Nuovo, Maurizio Palesi, Davide Patti, and Gianmarco 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.

Vincenzo Catania, Gianmarco De Francisci Morales, Alessandro G. Di Nuovo, Maurizio Palesi, and Davide Patti. "High Performance Computing for Embed-

ded System Design: A Case Study". In *DSD '08: 11th Conference on Digital System Design Architectures, Methods and Tools*, pp. 656–659, 2008.