Djellel Difallah

CONTACT INFORMATION

Phone +41 76 671 2383
E-mail: djellel@nyu.edu
Webpage: https://dedcode.github.io

RESEARCH INTERESTS

Data Science, Human Computation, Collaborative Knowledge Bases, Data Management, Information Extraction.

EDUCATION

University of Fribourg, Fribourg, Switzerland 2011 to 2015
Ph.D. Computer Science (Best CS Thesis Award) – 06/2015
- Dissertation title: "Quality of Service in Crowd-Powered Systems".
- Adviser: Prof. Philippe Cudré-Mauroux.

University of Louisiana, Lafayette, USA 2009 to 2011
M.S. Computer Science (with Honors) – May 2011
- Project: "Frequent Association Action Rules Mining Using FP-Trees".
- Adviser: Prof. Vijay V. Raghavan.

University of Sciences and Technology HB, Algiers, Algeria 1999 to 2004
Dipl.-Ing. Informatics – December 2004
- Track: Information Systems.
- Project: "Modeling the information system of Air-Algeria medical centers".

PROFESSIONAL EMPLOYMENT

Assistant Professor of Computer Science, New York University, Abu Dhabi Since 09/2020
Division of Science.

Research Scientist, Wikimedia Foundation 10/2019 to 08/2020
Research Team
- Link recommendation, sock-puppet detection, knowledge gaps.

Adjunct Assistant Professor, NYU Spring 2019
Leonard N. Stern School of Business
- Teaching data science to graduate business students (MBAs).

Postdoctoral Fellow, NYU 06/2017 to 04/2019
Center for Data Science
- **Population Estimation:** Develop statistical techniques for crowd size estimation in online communities (using Amazon MTurk surveys), and urban areas (using Foursquare data).
- **Edit Patterns Analysis in Knowledge Graphs:** Analysis of Wikidata editors engagement patterns. Research problems include lifetime prediction, edit patterns classification, and Wikidata class completeness estimation.
- **Ontology-based Data Integration:** Data integration and schema matching with Wikidata.

Senior Research Scientist, University of Fribourg (Switzerland) 09/2015 to 05/2017
eXascale Infolab
- **Ontology-based methods in NLP:** Develop methods for NLP tasks: entity linking, type inference and co-reference resolution, using DBPedia and Word2Vec for entity embedding, with applications to job information retrieval.
- **Academic Activities:** Teaching graduate courses (Big Data Infrastructures and Data Science Seminar); involved in writing two successfully funded projects ($\approx$ \$3 million); and co-supervised seven PhD/Master/Bachelor students.

Research Assistant, University of Fribourg (Switzerland) 05/2011 to 06/2015
eXascale Infolab

- **Crowd-powered Systems:** Research in system architectures and methods for human task scheduling, pricing, and recommendation.
- **Resource Management in Big Data Systems:** Develop modules for Apache Hadoop including job scheduling and block placement strategies – in collaboration with Microsoft CISL and Verisign.
- **Array Disk Storage Manager:** Developed a system for efficiently storing and retrieving multidimensional arrays on disk – in collaboration with SciDB.
- **Anomaly Detection in Big Data Systems:** Developed a stream processing system for anomaly detection in Water Distribution Networks – in collaboration with IBM Smart Cities.
- **Benchmarking:** Developed benchmarking software for RDBMSs and Array-DBMSs.

Intern, Microsoft Research, Mountain View, CA May to August 2013
Cloud and Information Services Lab

- Project: "Reservation-based scheduling with Hadoop YARN".
- Intern Supervisor: Carlo Curino.

Student Intern, Google (Summer of Code Program) May to August 2010
Drizzle DBMS

- Project: "A query cache plugin for Drizzle DBMS based on Memcached".
- Project Supervisor: Toru Maesaka.

Information Management Engineer, Schlumberger, Algiers, Agleria 2006 to 2009
Schlumberger Information Solutions

- Data management in the oil field industry. Implement ETL processes that ingest data originating from remote field locations into a centralized data warehouse.
- Perform Oracle Database Administration.
- Write and optimize SQL queries.
- Supervisors: Toufik Bessadat, Mounir Ferroudj.

Software Developer, EEPAD Internet Services Provider 2005 to 2006
Authentication and Billing Department

- Integration of authentication systems.
- Develop a network monitoring system using SNMP.
- Supervisor: Yacine Rezgui.

IT Manager, ZF Algeria 2004

TEACHING EXPERIENCE

Stern School of Business, NYU, New York, USA
Instructor Spring 2019

- Data Mining for Business Analytics (Managerial/Graduate section).
- Topics: introduction to data science with business use-cases; hands-on with Weka.

University of Fribourg, Fribourg, Switzerland
Co-Instructor Fall 2016

- Big Data Infrastructures (Graduate)
  - Co-Instructor: Prof. Philippe Cudré-Mauroux.
  - Topics: Introduction to DBMS/SQL, Distributed Data Management, NOSQL, Graph Databases, Hadoop, Spark.

Co-Instructor Fall 2015

- Data Science Seminar (Graduate)
  - Co-Instructor: Dr. Mourad Khayati.
  - Topics: clustering, compression and similarity techniques used for time series data and graphs.

Teaching Assistant Fall semesters from 2011 to 2014

- Social Computing (Graduate)
  - Main instructors: Dr. Gianluca Demartini and Prof. Philippe Cudré-Mauroux.
  - Topics: Semantic Web, Crowdsourcing, and Graphs.

Lucerne University of Applied Sciences, Lucerne, Switzerland

Guest Lecture Speaker Summer 2016

- Introduction to Big Data.

International Institute of Management in Technology, Fribourg, Switzerland

Guest Lecture Speaker Summer 2016

- Big Data hands-on for executives.

GRANTS AND AWARDS

Senior Staff – European Project H2020 2017

- Project: "FashionBrain: Understanding Europe's Fashion Data Universe".
- Project track: "Big Data PPP: Cross-sectorial and Cross-lingual Data Integration and Experimentation".
- Grant value: 2.9M EUR total; University of Fribourg's share: 690'000 EUR.

Senior Staff – Swiss Commission of Technology and Innovation 2016

- Project: "Query Expansion: Deep Learning and Crowdsourcing".
- Grant value: 210'000 CHF.

Best Computer Science thesis award – University of Fribourg February 2016

- Project: "Measuring Human Behaviour in Search Evaluation Micro Tasks".
- Grant value: 3'000 CHF.

Research Visit Grant – European Science Foundation Spring 2015

- Project: "Measuring Human Behaviour in Search Evaluation Micro Tasks".
- Host: Dr. Gianluca Demartini, University of Sheffield, UK.
- Grant value: 1'350 EUR.

Travel Grant – HCOMP Doctoral Consortium October 2014

- Project: "Scalable Human-based Computation".
- Grant value: 700 USD.

University of Louisiana at Lafayette Honors, for maintaining a GPA of 4.0 2011

Fulbright Exchange Student Grant 2009-2011

INVITED TALKS

- Djellel Difallah. "Crowd Size Estimation" (Seminar)
  Moore-Sloan Data Science Summit. Utah, USA September 2018
- Ujwal Gadiraju, Gianluca Demartini, Djellel Difallah and Michele Catasta. "Using Crowdsourcing Effectively for Social Media Research" (Tutorial)
  ICWSM'16 Cologne and WebSci'16 Hannover. Germany May 2016
- Djellel Difallah. "Quality and Performance Optimizations in Microtask Crowdsourcing" (Seminar)
  MIT CSAIL, Boston, MA and CMU, Pittsburgh, PA November 2014
- Djellel Difallah. "Extending The OLTP-Bench Framework for Big Data Systems" (Workshop)
  Workshop on Big Data Benchmarking, Potsdam, Germany August 2014
- Djellel Difallah. "Resource Management and Planning for Hadoop" (Internal Talk)
  Microsoft Research. Mountain View, CA August 2013
- Djellel Difallah. "Mechanical Cheat: Spamming Schemes and Adversarial Techniques on Crowdsourcing Platforms" (Workshop)
  CrowdSearch 2012 workshop at WWW, Lyon, France April 2012

Professional Service

Program Committee Member

- *25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2019)*
- *7th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2019)*
- *28th International World Wide Web Conference (WWW 2019)*
- *27th ACM Conference on Information and Knowledge Management (CIKM 2018)*
- *IEEE International Conference on Data Mining (ICDM 2018)*
- *24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2018)*
- *12th International AAAI Conference On WEB and Social Media (ICWSM 2018)*
- *27th International World Wide Web Conference (WWW 2018)*
- *23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2017)*
- *26th ACM Conference on Information and Knowledge Management (CIKM 2017)*
- *26th International World Wide Web Conference (WWW 2017)*
- *16th International Conference on Web Engineering (ICWE 2016)*

External Reviewer

- IEEE International Conference on Data Engineering – ICDE (2016)
- ACM Special Interest Group on Management of Data – SIGMOD (2016)
- ACM Conference on Web Science – WebSci (2016)
- ACM Symposium on Cloud Computing – SoCC (2016)
- VLDB (2014, 2016)
- EDBT (2013)
- AAAI (2014, 2015)
- IEEE BigData (2013)

Open Source Service

OLTP-Bench (Lead Contributor 2012-2015)

- Database Benchmarking Framework
- https://github.com/oltpbenchmark

OpenTurk (Lead Contributor 2013-2014)

- Chrome extension for Mturk HIT notification
- https://github.com/openturk

Apache Hadoop YARN (Contributor 2013)

- Hadoop YARN Admission Control/Planner
- Code Repository: https://issues.apache.org/jira/browse/YARN-1051
- Various bug fixes in the scheduler

SciDB (Contributor 2011-2013)

- Array Database Management System
- Implementation of a science benchmark (SSDB)

Drizzle (Contributor 2010)

- Database Management System with a micro-kernel architecture
- Code Repository: https://code.launchpad.net/~dedzone/drizzle/query-cache-soc

Apache Mahout (Contributor 2010)

- Distributed machine-learning algorithms
- Bug fixes in K-Means implementation

Advising and Mentoring

- **Alexander Striffeler**, Master Student, Computer Science 2015
  Thesis topic: "Benchmarking Big Data Infrastructures"
  Co-advised by: Philippe Cudré-Mauroux and Mourad Khayati
- **Marwa Bouzeyane**, Master Student, Computer Science 2015
  Thesis topic: "Grammatical Errors Detection Based On N-gram Analysis"
  Co-advised by: Philippe Cudré-Mauroux and Mourad Khayati

- **Stefan Nüesch**, Master Student, Computer Science 2014
  Thesis topic: "Real-Time Anomaly Detection in Water Distribution Networks using Spark Streaming"
  Co-advised by: Philippe Cudré-Mauroux
- **Phokham Nonava**, Master Student, Computer Science 2014
  Thesis topic: "HDFS Blocks Placement Strategy"
  Co-advised by: Philippe Cudré-Mauroux, Benoit Perroud and Martin Grund
- **Simpal Kumar**, Master Student, Computer Science 2014
  Thesis topic: "Real Time Data Analysis for Water Distribution Network using Storm"
  Co-advised by: Philippe Cudré-Mauroux
- **Dani Rotzetter**, Master Student, Computer Science 2014
  Thesis topic: "Crowd-Flow Designer: An Open-Source Toolkit to Design and Run Complex Crowd-Sourced Tasks"
  Co-advised by: Philippe Cudré-Mauroux and Gianluca Demartini
- **Victor Felder**, Bachelor Student, Computer Science 2013
  Thesis topic: "Openturk: An Implementation of the Pick-A-Crowd Architecture"

COMMUNITY SERVICES

- **Software Carpentry Student Support**. Introduction to Git, Unix shell, R, Python for non-programmers at NYU (2018).
- **Job-Info Day Instructor**. Introducing the computer science track to high school students in Fribourg (1 day event – from 2013 to 2017).
- **Cyber-Camp Instructor**. Summer camp for high school students in Fribourg (1 week event – from 2012 to 2014).

JOURNAL PUBLICATIONS

D. Difallah, A. Checco, G. Demartini, and P. Cudré-Mauroux. Deadline-aware fair scheduling for multi-tenant crowd-powered systems. *ACM Trans. Social Computing*, 2(1), 2019

C. Sarasua, A. Checco, G. Demartini, **D. Difallah**, M. Feldman, and L. Pintscher. The evolution of power and standard wikidata editors: Comparing editing behavior over time to predict lifespan and volume of edits. *Computer Supported Cooperative Work (CSCW)*, Dec 2018

G. Demartini, **D. Difallah**, and P. Cudré-Mauroux. Large-scale linked data integration using probabilistic reasoning and crowdsourcing. *VLDB J.*, 22(5):665–687, 2013

D. Difallah, P. Cudré-Mauroux, and S. A. McKenna. Scalable anomaly detection for smart city infrastructure networks. *IEEE Internet Computing*, 17(6):39–47, 2013

CONFERENCE PUBLICATIONS

M. Luggen, **D. Difallah**, C. Sarasua, G. Demartini, and P. Cudré-Mauroux. Non-parametric class completeness estimators for collaborative knowledge graphs – the case of wikidata. 2019

D. Difallah, E. Filatova, and P. Ipeirotis. Demographics and dynamics of mechanical turk workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pages 135–143, New York, NY, USA, 2018. ACM

J. Plu, R. Prokofyev, T. Tonon, P. Cudré-Mauroux, **D. Difallah**, R. Troncy, and G. Rizzo. Sanaphor++: Combining deep neural networks with semantics for coreference resolution. LREC, 2018

R. Prokofyev, M. Luggen, **D. Difallah**, and P. Cudré-Mauroux. Swisslink: High-precision, context-free entity linking exploiting unambiguous labels. In *Proceedings of the 13th International Conference on Semantic Systems*. ACM, 2017

D. Difallah, G. Demartini, and P. Cudré-Mauroux. Scheduling human intelligence tasks in multi-tenant crowd-powered systems. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 855–865, 2016

A. Tonon, V. Felder, **D. Difallah**, and P. Cudré-Mauroux. Voldemortkg: Mapping schema.org and web entities to linked open data. In *The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016, Proceedings, Part II*, pages 220–228, 2016

R. Prokofyev, A. Tonon, M. Luggen, L. Vouilloz, **D. Difallah**, and P. Cudré-Mauroux. SANAPHOR: ontology-based coreference resolution. In *The Semantic Web - ISWC 2015 - 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part I*, pages 458–473, 2015

D. Difallah, M. Catasta, G. Demartini, P. G. Ipeirotis, and P. Cudré-Mauroux. The dynamics of micro-task crowdsourcing: The case of amazon mturk. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*, pages 238–247, 2015

M. Catasta, A. Tonon, **D. Difallah**, G. Demartini, K. Aberer, and P. Cudré-Mauroux. Hippocampus: answering memory queries using transactive search. In *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, Companion Volume*, pages 535–540, 2014

M. Catasta, A. Tonon, **D. Difallah**, G. Demartini, K. Aberer, and P. Cudré-Mauroux. Transactivedb: Tapping into collective human memories. *Proceedings of the VLDB Endowment*, 7(14):1977–1980, 2014

C. Curino, **D. Difallah**, C. Douglas, S. Krishnan, R. Ramakrishnan, and S. Rao. Reservation-based scheduling: If you're late don't blame us! In *Proceedings of the ACM Symposium on Cloud Computing, Seattle, WA, USA, November 03 - 05, 2014*, pages 2:1–2:14, 2014

D. Difallah, M. Catasta, G. Demartini, and P. Cudré-Mauroux. Scaling-up the crowd: Micro-task pricing schemes for worker retention and latency improvement. In *Proceedings of the Seconf AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2014, November 2-4, 2014, Pittsburgh, Pennsylvania, USA*, 2014

D. Difallah, A. Pavlo, C. Curino, and P. Cudré-Mauroux. Oltp-bench: An extensible testbed for benchmarking relational databases. *Proceedings of the VLDB Endowment*, 7(4):277–288, 2013

D. Difallah, G. Demartini, and P. Cudré-Mauroux. Pick-a-crowd: tell me what you like, and i'll tell you what to do. A crowdsourcing platform for personalized human intelligence task assignment based on social networks. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*, pages 367–374, 2013

G. Demartini, **D. Difallah**, and P. Cudré-Mauroux. Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012*, pages 469–478, 2012

BOOK

G. Demartini, **D. Difallah**, U. Gadiraju, M. Catasta, et al. An introduction to hybrid human-machine information systems. *Foundations and Trends® in Web Science*, 7(1):1–87, 2017

BOOK CHAPTERS

D. Difallah, P. Cudré-Mauroux, S. McKenna, and D. Fasel. Skalierbar anomalien erkennen für smart city infrastrukturen. In *Big Data*, pages 289–299. Springer, 2016

WORKSHOPS AND DEMONSTRATIONS

D. V. Aken, **D. Difallah**, A. Pavlo, C. Curino, and P. Cudré-Mauroux. Benchpress: Dynamic workload control in the oltp-bench testbed. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Victoria, Australia, May 31 - June 4, 2015*, pages 1069–1073, 2015

C. Curino, **D. Difallah**, A. Pavlo, and P. Cudré-Mauroux. Benchmarking oltp/web databases in the cloud: the oltp-bench framework. In *Proceedings of the Fourth International Workshop on Cloud Data Management, CloudDB 2012, Maui, HI, USA, October 29, 2012*, pages 17–20, 2012

D. Difallah, G. Demartini, and P. Cudré-Mauroux. Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms. In *Proceedings of the First International Workshop on Crowdsourcing Web Search, Lyon, France, April 17, 2012*, pages 26–30, 2012

P. Cudré-Mauroux, G. Demartini, **D. Difallah**, A. Mostafa, V. Russo, and M. Thomas. A Demonstration of DNS3: a Semantic-Aware DNS Service. Citeseer, 2011

D. Difallah, R. G. Benton, V. V. Raghavan, and T. Johnsten. FAARM: frequent association action rules mining using fp-tree. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, Vancouver, BC, Canada, December 11, 2011*, pages 398–404, 2011

D. Difallah, P. Cudré-Mauroux, I. Stratos, and Y. Zhang. SSDB Benchmark: Implementations and Results. XLDB, 2011