Michele (pirroh) Catasta

About Me

Head of Applied Research at Google Labs
working on *AI applied to Source Code, Large Language Models*

Personal page: `https://pirroh.fyi`
LinkedIn profile: `https://linkedin.com/in/pirroh`
Twitter profile: `https://twitter.com/pirroh`
GitHub profile: `https://github.com/pirroh`
Google Scholar: `https://scholar.google.com/citations?user=ZxfOSWMAAAAJ&hl=en&oi=ao`

email

✉ `pirroh@cs.stanford.edu` (personal) ✉ `pirroh@google.com` (work)

Areas of Expertise

AI4Code, Machine Learning, Data Science, Information Retrieval

Education

Stanford University

Postdoc and Instructor in Machine Learning *2017 – 2019*

- Advisor: Prof. Jure Leskovec
- Affiliated with *Statistical Machine Learning Group*, SNAP and InfoLab

EPFL, Switzerland

Ph.D. in Computer Science *2010 – 2015*

- Advisors: Prof. Karl Aberer and Prof. Philippe Cudre-Mauroux
- Thesis: *MEM0R1ES: Memory-based Information Systems*

National University of Ireland, Galway

M.Sc. with Research Honors in Applied Computing and IT *2007 – 2009*

- Advisors: Prof. Stefan Decker and Dr. Giovanni Tummarello
- Thesis: *Scalability Solutions for the Web of Data*
- Full tuition scolarship and Research Assistantship contract

Experience

Irregular Expressions

Limited Partner and Advisor *2022 – Present*

IrregEx is a group of seasoned SWE and product leaders advising and investing in early-stage startups focused on developer tools.

X, the moonshot factory [formerly Google X] (Mountain View, CA, USA)

Head of Applied Research *2021 – 2022*

AI applied to Source Code, Large Language Models

Stanford University (Stanford, CA, USA)

Research Scientist and Instructor *2020*

Contributed to several projects: Graph Neural Networks and Transformer architectures applied to relational data (with a focus on Source Code), Open Graph Benchmark, recommender system for Wikipedia editors, resource allocation on crowdsourcing platforms (IARPA HFC), machine common sense (DARPA MCS).

EPFL (Lausanne, Switzerland)

Scientific Collaborator and Lecturer *2015 – 2017*

Created and taught the first edition of Applied Data Analysis (`https://ada.epfl.ch`), a brand new course in the EPFL Data Science specialization. Managed several research projects with a focus on machine learning, recommender systems, data mining, etc.

MIT Media Lab (Cambridge, MA, USA)

Visiting PhD in Prof. Sandy Pentland's group *Summer 2014*

Implemented a gesture-based authentication prototype on iOS, used by the Human Dynamics lab.

Google (Mountain View, CA, USA)

Software Engineer Intern, Infrastructure Storage *Summer 2011*

Given the resource-consumption information and the SLAs of the storage layer, I worked on a compiler that translates high-level description of SLOs into low-level monitoring code (using Python) and on prediction models for timeseries (using Dremel).

Yahoo Research (Barcelona, Spain)

Student Researcher in Prof. Ricardo Baeza-Yates' group *Summer 2008*

Developed a Big Data processing pipeline (using Hadoop and Pig) to extract metadata (such as `schema.org` tags) from the Yahoo Web crawl.

DERI (Galway, Ireland)

Research Assistant *2007 – 2009*

Co-founder of Sindice.com, the largest Semantic Web search engine at the time. I designed and developed several key components of the backend, with a focus on the scalability challenges (using MapReduce, Hadoop, Lucene, etc.) The core technologies developed for Sindice.com evolved into an investigative intelligence platform (`https://siren.io`) which—to date—has secured $15M+ in funding.

Grants and Awards

- ◇ Granted $5,000 in Google Cloud Platform credits for the research project Academic Diaspora, *2020*
- ◇ Nominated **Google Cloud Faculty Expert**, *2020*
- ◇ Received an Education Grant for $195,000 in Google Cloud Platform credits to support the course CS224W: Machine Learning with Graphs, *2019*
- ◇ Granted $15,000 in Amazon Web Services credits for the research project WebGenome, *2018*
- ◇ Visiting Scholar grant from the Information School of the University of Sheffield, *2016*
- ◇ Google Student grant for the World Wide Web conference, Firenze, Italy, *2015*
- ◇ Short Visit grant for "Evaluating Information Access Systems", University of Sheffield, *2015*
- ◇ Invited to the Global Young Scientists Summit @one-north, Singapore, *2015*
- ◇ Lead author of the CROSS grant (60,000 CHF), multidisciplinary collab. with UNIL, *2014*
- ◇ 3rd prize at the Semantic Web Challenge, ISWC, *2013*
- ◇ Lead author of the MemorySense grant ($150,000), collab. with Samsung Research USA, *2013*
- ◇ Outstanding Teaching Assistant Award, EPFL, *2012*
- ◇ Hypothes.is Reputation Fellow, *2012*
- ◇ 3rd prize at the Semantic Web Challenge, ISWC, *2009*
- ◇ 2nd prize at the Asian Semantic Web Conference, Industrial Track, *2009*
- ◇ 1st prize at the European Semantic Web Technology Conference, Business Idea Contest, *2008*
- ◇ Saltlux Best Paper Award for the Sindice journal paper, *2007*

Teaching Activities

- ◇ **Advisor** for CS329S: Machine Learning Systems Design at Stanford University, *2021*
- ◇ **Co-instructor** for CS224W: Machine Learning with Graphs (≈200 students) and CS246: Mining Massive Data Sets (≈300 students) at Stanford University, *2018 – 2020*
- ◇ Mentor at Lead The Future, a network to support Italian excellence in STEM, *2020 – Present*
- ◇ **Instructor** for CS341: Project in Mining Massive Data Sets at Stanford University, *2019*
- ◇ Creator and first instructor of Applied Data Analysis (≈250 students), a brand new course in Data Science at EPFL, **second largest course offered by the CS department**, *2016 – 2017*
- ◇ Core member of the committee which developed the Data Science MSc at EPFL, *2015 – 2017*
- ◇ Lecturer for the course Distributed Information Systems at EPFL, *2015 – 2016*
- ◇ Lead TA for the course Distributed Information Systems at EPFL, *2011 – 2014*
- ◇ TA for the course Introduction to Database Systems at EPFL, *2012*
- ◇ Co-founder of "Hackers at EPFL", organizing hackathons, tech talks, etc., *2012 – 2015*

Academic Activities

- ◇ Co-organizer of the Mining and Learning with Graphs workshop at KDD, *2017*
- ◇ Co-organizer of the Dagstuhl Seminar Crowdsourcing Research - Transcending Disciplinary Boundaries, *2016*

◇ **Senior PC member** of CIKM 2020, CIKM 2021
◇ **PC member** of SocInfo 2012, CIKM 2014, ESWC 2015, ICWE 2016, ESWC 2016, ISWC 2016 Doctoral Consortium, HCOMP 2016, WWW 2017 (and posters), Wiki Workshop 2017, ESWC 2017, KDD 2017, WebSci 2017, ISWC 2017 Doctoral Consortium, CIKM 2017, AAAI 2018, WSDM 2018, WWW 2018 (and posters), Wiki Workshop 2018, WebSci 2018, KDD 2018 Applied Data Science, CIKM 2018, ICDM 2018, HumL at WWW 2018 and ISWC 2018, CrowdBias Workshop at HCOMP 2018, Euro CSS Symposium 2018, WSDM 2019, Wiki Workshop 2019, KDD 2019, WWW 2020, ECML-PKDD 2020, GRL+ Workshop @ ICML 2020, ICML 2020, NeurIPS 2020, ICLR 2021, AAAI 2021, ICML 2021, NeurIPS 2021, ICLR 2022, ICML 2022.
◇ **Reviewer** for ACM Transactions on Social Computing, IEEE Transactions on Knowledge and Data Engineering, EPJ Data Science, Nature Scientific Data, Harvard Data Science Review.

Invited Talks

◇ *AI meets Source Code: status quo and outlooks*, Synapse AI Symposium, *2022*
◇ *AI meets Source Code: status quo and outlooks*, EPFL, hosted by Prof. Robert West, *2022*
◇ *AI meets Source Code: status quo and outlooks*, University of Fribourg, hosted by Prof. Philippe Cudre-Mauroux, *2022*
◇ Moderator of *Panel: Big Data and AI in the Classroom with Google Cloud*, Google Faculty Institute, *2020*
◇ *Advancements in Graph Deep Learning*, Graph The Planet, Bloomberg SF, *2020*
◇ *Structuring Wikipedia Articles with Section Recommendations*, University of Fribourg, hosted by Prof. Philippe Cudre-Mauroux, *2018*
◇ *Recommender Systems in the Wild*, Papers We Love, Square SF, *2017*
◇ *Big Data Mining*, TUI Strategy Workshop in AI and ML, Palma de Mallorca, *2017*
◇ *Recommender Systems in the Wild*, Università Svizzera Italiana, hosted by Prof. Crestani, *2017*
◇ Lecturer at the CUSO Winter School "*Data Science in Information Society: From Data Acquisition to Data Analysis*", *2017*
◇ *Entity Typing on the Web*, Sapienza University of Rome, hosted by Prof. Roberto Navigli, *2016*
◇ *MEM0R1ES: Memory-based Information Systems*, University of Leeds, hosted by Prof. Ruddle, *2016*
◇ *MEM0R1ES: Memory-based Information Systems*, University of Sheffield, hosted by Prof. Gianluca Demartini, *2016*
◇ *Entity Typing on the Web*, Yahoo Labs London, hosted by Dr. Peter Mika, *2015*
◇ Invited lecture for the *Technology Ventures in IC* course at EPFL, hosted by Prof. Bugnion, *2015*
◇ *MEmoIt: don't write your diary, sense it*, DBTA Workshop on "Life Logging and Long-Term Digital Preservation" at University of Lugano, *2015*
◇ *MemorySense: let your smartphone figure out the memorable moments of your life*, Samsung Research America, *2014*

Selected Publications

A Chowdhery, S Narang, J Devlin, M Bosma, G Mishra et al. "PaLM: Scaling Language Modeling with Pathways", *arXiv:2204.02311*

D Zugner, T Kirschstein, M Catasta, J Leskovec, S Gunnemann, "Language-Agnostic Representation Learning of Source Code from Structure and Context", *ICLR 2021*

T Vaucher, A Spitz, M Catasta, R West, "A Decade in Quotes: Minimally Supervised Quotation Attribution in Massive News Corpora", *WSDM 2021*, **Oral**

W Hu, M Fey, M Zitnik, Y Dong, H Ren, B Liu, M Catasta, J Leskovec "Open graph benchmark: Datasets for machine learning on graphs", *NeurIPS 2020*, **Spotlight**

T Piccardi, M Catasta, L Zia, R West, "Structuring Wikipedia Articles with Section Recommendations", *SIGIR 2018*

J Rappaz, M Catasta, R West, K Aberer, "Latent Structure in Collaboration: the case of Reddit r/place", *ICWSM 2018*

J Rappaz, ML Vladarean, J McAuley, M Catasta, "Bartering books to beers: A recommender system for exchange platforms", *WSDM 2017*

JE Ranvier, M Catasta, M Vasirani, K Aberer, "RoutineSense: A Mobile Sensing Framework for the Reconstruction of User Routines", *Mobiquitous 2015*, **Best Paper Award**

DE Difallah, M Catasta, G Demartini, P Ipeirotis, P Cudre-Mauroux, "The Dynamics of Micro-Task Crowdsourcing – The Case of Amazon MTurk", *WWW 2015*

DE Difallah, M Catasta, G Demartini, P Cudre-Mauroux, "Scaling-up the Crowd: Micro-Task Pricing Schemes for Worker Retention and Latency Improvement", *HCOMP 2014*

M Catasta, A Tonon, DE Difallah, G Demartini, K Aberer, P Cudre-Mauroux, "TransactiveDB: Tapping into Collective Human Memories", *VLDB 2014*

K Aberer, M Catasta, G Christodoulou, I Gavrilovic, F Hrisafov, M Monney, A Ouaazki, B Perovic, H Radu, JE Ranvier, M Vasirani, Z Yan, "Memo-it: Don't write your diary, Sense it", *UbiComp 2014*

M Catasta, A Tonon, DE Difallah, G Demartini, K Aberer, P Cudre-Mauroux, "Hippocampus: Answering Memory Queries using Transactive Search", *WWW 2014*

K Aberer, M Catasta, H Radu, JE Ranvier, M Vasirani, Z Yan, "MemorySense: Reconstructing and Ranking User Memories on Mobile Devices", *PerCom 2014*

A Tonon, M Catasta, G Demartini, P Cudre-Mauroux, K Aberer, "TRank: Ranking Entity Types Using the Web of Data", *ISWC 2013*, **Best Paper Award nominee**

LH Tran, M Catasta, LK McDowell, K Aberer, "Next Place Prediction using Mobile Data", *Mobile Data Challenge (by Nokia) Workshop 2012*

R Delbru, N Toupikov, M Catasta, G Tummarello, "A Node Indexing Scheme for Web Entity Retrieval", *ESWC 2010*

R Delbru, N Toupikov, M Catasta, G Tummarello, S Decker, "Hierarchical Link Analysis for Ranking Web Data", *ESWC 2010*

Journal Papers

T Sinha, M Kapur, R West, M Catasta, M Hauswirth, D Trninic "Differential Benefits of Explicit Failure-driven and Success-driven Scaffolding in Problem-Solving Prior to Instruction", *Journal of Educational Psychology*, 2020

A Tonon, M Catasta, R Prokofyev, G Demartini, K Aberer, P Cudre-Mauroux "Contextualized ranking of entity types based on knowledge graphs", *Journal of Web Semantics*, 2016

M Catasta, A Tonon, V Pasquier, G Demartini, P Cudre-Mauroux, K Aberer, "B-hist: Entity-Centric Search over Personal Web Browsing History", *Journal of Web Semantics*, 2014

G Tummarello, R Cyganiak, M Catasta, S Danielczyk, R Delbru, S Decker, "Sig.ma: Live Views on the Web of Data", *Journal of Web Semantics*, 2010, **Top-25 Most Cited article of JoWS**

E Oren, R Delbru, M Catasta, R Cyganiak, H Stenzhorn, G Tummarello, "Sindice.com: A Document-oriented Lookup Index for Open Linked Data", *International Journal of Metadata, Semantic and Ontologies*, 2008

Book Chapters

JE Ranvier, M Catasta, I Gavrilovic, G Christodoulou, H Radu, K Aberer, "MEmoIt: From Lifelogging Application to Research Platform", Book chapter in *Mobile Application Development, Usability, and Security*, IGI Global, 2016

R Delbru, N Toupikov, M Catasta, R Fuller, G Tummarello, "SIREn: Efficient Search on Semi-Structured Documents", Book chapter in *Lucene in Action 2nd Edition* (In Action series), Manning Publications Co., 2009

M Catasta, R Delbru, N Toupikov, G Tummarello, "Managing Terabytes of Web Semantics Data", Book chapter in *Semantic Web Information Management*, Springer, 2009

Book

G Demartini, DE Difallah, U Gadiraju, M Catasta, "An Introduction to Hybrid Human-Machine Information Systems", Book in *Foundations and Trends in Web Science*, NOW, 2017

Tutorials

- "It's Getting Crowded!": How To Use Crowdsourcing Effectively for Web Science Research, `https://itsgettingcrowded.wordpress.com`, *WebSci 2016*
- Using Crowdsourcing Effectively for Social Media Research, *ICWSM 2016*