1st International Workshop on Evaluating Progress in Artificial Intelligence - EPAI 2020
In conjunction with the 24th European Conference on Artificial Intelligence - ECAI 2020
*Santiago de Compostela, Spain*

# The Scientometrics of AI Benchmarks: Unveiling the Underlying Mechanics of AI Research

**Pablo Barredo**[1] and **José Hernández-Orallo**[2]
and **Fernando Martínez-Plumed**[2,3] and **Seán Ó hÉigeartaigh**[4]

**Abstract**

The widespread use of experimental benchmarks in AI research has created new competition and collaboration dynamics that are still poorly understood. In this paper we provide an innovative methodology to explore this dynamics and analyse the way different entrants in these competitions, from academia to tech giants, behave and react depending on their own or others' achievements. We perform an analysis of over twenty popular benchmarks in AI, linking their underlying research papers. We identify links between researchers and institutions (i.e., communities) beyond the standard co-authorship relations, and we explore a series of hypotheses about their behaviour as well as some aggregated results in terms of activity, breakthroughs and efficiency. As a result, we detect and characterise the dynamics of research communities at different levels of abstraction, including organisation, affiliation, trajectories, results and activity.

**Submission under review**

[1] Universidad de Oviedo, email: UO237136@uniovi.es
[2] Universitat Politècnica de València, email: jorallo@upv.es, fmartinez@dsic.upv.es
[3] JRC, European Commission, email: fernando.martinez-plumed@ec.europa.eu
[4] University of Cambridge, email: so348@cam.ac.uk

## ACKNOWLEDGEMENTS

This material is based upon work supported by the EU (FEDER), and the Spanish MINECO under grant RTI2018-094403-B-C3, the Generalitat Valenciana PROMETEO/2019/098. F. Martínez-Plumed acknowledges funding of the AI-Watch project by DG CONNECT and DG JRC of the European Commission. J. Hernández-Orallo and S. Ó hÉigeartaigh are also funded by an FLI grant RFP2-152. We thank the anonymous comments received by the EPAI committee.

## REFERENCES

[1] Jose M Alonso, Ciro Castiello, and Corrado Mencar, 'A bibliometric analysis of the explainable artificial intelligence research field', in *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp. 3–15. Springer, (2018).

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele, '2D human pose estimation: New benchmark and state of the art analysis', in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2014).

[3] Jay Bhattacharya and Mikko Packalen, 'Stagnation and scientific incentives', Technical report, National Bureau of Economic Research, (2020).

[4] Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna, 'Findings of the 2014 workshop on statistical machine translation', in *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 12–58, Baltimore, Maryland, USA, (June 2014). Association for Computational Linguistics.

[5] Carlo Bonferroni, 'Teoria statistica delle classi e calcolo delle probabilita', *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, **8**, 3–62, (1936).

[6] M. Campbell, A. J. Hoane, and F. Hsu, 'Deep Blue', *Artificial Intelligence*, **134**(1-2), 57 – 83, (2002).

[7] Stephen Cave and Seán S ÓhÉigeartaigh, 'An AI race for strategic advantage: rhetoric and risks', in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 36–40, (2018).

[8] Aaron Clauset, M. E. J. Newman, and Cristopher Moore, 'Finding community structure in very large networks', *Physical Review E*, **70**(6), (12 2004).

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, 'The cityscapes dataset for semantic urban scene understanding', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, (2016).

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, 'Imagenet: A large-scale hierarchical image database', in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, (2009).

[11] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, 'The pascal visual object classes challenge: A retrospective', *International Journal of Computer Vision*, **111**(1), 98–136, (January 2015).

[12] Ethan Fast and Eric Horvitz, 'Long-term trends in the public perception of artificial intelligence.', in *AAAI*, pp. 963–969, (2017).

[13] David A Ferrucci, 'Introduction to "This is Watson"', *IBM Journal of Research and Development*, **56**(3.4), 1–1, (2012).

[14] Fang Gao, Xiaofeng Jia, Zhiyun Zhao, Chih-Cheng Chen, Feng Xu, Zhe Geng, and Xiaotong Song, 'Bibliometric analysis on tendency and topics of artificial intelligence over last decade', *Microsystem Technologies*, 1–13, (2019).

[15] Danny Hernandez and Tom B Brown, 'Measuring the algorithmic efficiency of neural networks', *arXiv preprint arXiv:2005.04305*, (2020).

[16] José Hernández-Orallo, Marco Baroni, Jordi Bieger, Nader Chmait, David L Dowe, Katja Hofmann, Fernando Martínez-Plumed, Claes Strannegård, and Kristinn R Thórisson, 'A new AI evaluation cosmos: Ready to play the game?', *AI Magazine*, **38**(3), (2017).

[17] Michael C Horowitz, Gregory C Allen, Elsa B Kania, and Paul Scharre, 'Strategic competition in an era of artificial intelligence', *Center for New American Security (Washington, DC: Center for New American Security, 2018)*, **8**, (2018).

[18] Brandon Houghton, Stephanie Milani, Nicholay Topin, William Guss, Katja Hofmann, Diego Perez-Liebana, Manuela Veloso, and Ruslan Salakhutdinov, 'Guaranteeing reproducibility in deep learning competitions', *arXiv preprint arXiv:2005.06041*, (2020).

[19] Konstantinos Stathoulopoulos Juan Mateos-Garcia, Joel Klinger and Russell Winch. A semantic analysis of the recent evolution of ai research, 2019.

[20] Alex Krizhevsky. Learning multiple layers of features from tiny images. `https://www.cs.toronto.edu/~kriz/cifar.html`, 2009.

[21] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre, 'Hmdb: a large video database for human motion recognition', in *2011 International Conference on Computer Vision*, pp. 2556–2563. IEEE, (2011).

[22] Roberta Kwok, 'Junior AI researchers are in demand by universities and industry', *Nature*, **568**(7752), 581–584, (2019).

[23] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. http://yann.lecun.com/exdb/mnist/, 2010.

[24] Kai-Fu Lee, *AI superpowers: China, Silicon Valley, and the new world order*, Houghton Mifflin Harcourt, 2018.

[25] Yufei Lei and Zhongbao Liu, 'The development of artificial intelligence: a bibliometric analysis, 2007-2016', in *Journal of Physics: Conference Series*, volume 1168, p. 022027. IOP Publishing, (2019).

[26] Loet Leydesdorff, *The challenge of scientometrics: The development, measurement, and self-organization of scientific communications*, Universal-Publishers, 2001.
[27] Loet Leydesdorff and Staša Milojević, 'Scientometrics', *arXiv preprint arXiv:1208.4566*, (2012).
[28] Wendy CY Li, Makoto Nirei, and Kazufumi Yamana, 'Value of data: there's no such thing as a free lunch in the digital economy', *US Bureau of Economic Analysis Working Paper, Washington, DC*, (2019).
[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, 'Microsoft coco: Common objects in context', in *European conference on computer vision*, pp. 740–755. Springer, (2014).
[30] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet, 'Are gans created equal? a large-scale study', in *Advances in neural information processing systems*, pp. 700–709, (2018).
[31] Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts, 'Learning word vectors for sentiment analysis', in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pp. 142–150. Association for Computational Linguistics, (2011).
[32] Marlos C Machado, Marc G Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling, 'Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents', *Journal of Artificial Intelligence Research*, **61**, 523–562, (2018).
[33] Fernando Martínez-Plumed, Shahar Avin, Miles Brundage, Allan Dafoe, Sean Ó hÉigeartaigh, and José Hernández-Orallo, 'Accounting for the neglected dimensions of ai progress', *arXiv preprint arXiv:1806.00610*, (2018).
[34] Fernando Martínez-Plumed, Bao Sheng Loe, Peter Flach, Seán O hEigeartaigh, Karina Vold, and José Hernández-Orallo, 'The facets of artificial intelligence: A framework to track the evolution of AI', *IJCAI*, (2018).
[35] F. Martínez-Plumed and J. Hernández-Orallo, 'Dual indicators to analyze ai benchmarks: Difficulty, discrimination, ability, and generality', *IEEE Transactions on Games*, **12**(2), 121–131, (2020).
[36] Peter Mattson, Christine Cheng, Cody Coleman, Greg Diamos, Paulius Micikevicius, David Patterson, Hanlin Tang, Gu-Yeon Wei, Peter Bailis, Victor Bittorf, et al., 'Mlperf training benchmark', *arXiv preprint arXiv:1910.01500*, (2019).
[37] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher, 'Pointer sentinel mixture models', *arXiv preprint arXiv:1609.07843*, (2016).
[38] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, 'Human-level control through deep reinforcement learning', *Nature*, **518**, 529–533, (2015).
[39] Jiqiang Niu, Wenwu Tang, Feng Xu, Xiaoyan Zhou, and Yanan Song, 'Global research on AI from 1990–2014: Spatially-explicit bibliometric analysis', *ISPRS Int. Journal of Geo-Information*, **5**(5), 66, (2016).
[40] Carlos Purves, Cătălina Cangea, and Petar Veličković, 'The playstation reinforcement learning environment (psxle)', *arXiv preprint arXiv:1912.06101*, (2019).
[41] Pranav Rajpurkar, Robin Jia, and Percy Liang, 'Know what you don't know: Unanswerable questions for squad', *arXiv preprint arXiv:1806.03822*, (2018).
[42] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang, 'Squad: 100,000+ questions for machine comprehension of text', *arXiv preprint arXiv:1606.05250*, (2016).
[43] Stephen A Rhoades, 'The Herfindahl-Hirschman index', *Fed. Res. Bull.*, **79**, 188, (1993).
[44] Erik F Sang and Fien De Meulder, 'Introduction to the conll-2003 shared task: Language-independent named entity recognition', *arXiv preprint cs/0306050*, (2003).
[45] David Schlangen, 'Language tasks and language games: On methodology in current natural language processing research', *arXiv preprint arXiv:1908.10747*, (2019).
[46] Yoav Shoham, 'Towards the AI index', *AI Magazine*, **38**(4), 71–77, (2017).
[47] David Silver et al., 'Mastering the game of Go with deep neural networks and tree search', *Nature*, **529**(7587), 484, (2016).
[48] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts, 'Recursive deep models for semantic compositionality over a sentiment treebank', in *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, (2013).
[49] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, 'Ucf101: A dataset of 101 human actions classes from videos in the wild', *arXiv preprint arXiv:1212.0402*, (2012).
[50] Bach Xuan Tran et al., 'Global evolution of research in artificial intelligence in health and medicine: A bibliometric study', *Journal of clinical medicine*, **8**(3), 360, (2019).
[51] Anthony Van Raan, 'The influence of international collaboration on the impact of research results: Some simple mathematical considerations concerning the role of self-citations', *Scientometrics*, **42**(3), 423–428, (1998).
[52] Oriol Vinyals et al., 'Starcraft II: A new challenge for reinforcement learning', *arXiv preprint arXiv:1708.04782*, (2017).
[53] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman, 'Superglue: A stickier benchmark for general-purpose language understanding systems', *arXiv preprint arXiv:1905.00537*, (2019).
[54] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi, 'Hellaswag: Can a machine really finish your sentence?', *arXiv preprint arXiv:1905.07830*, (2019).