

NeurIPS Datasets and Benchmarks 2021

Online
6 - 14 December 2021

Volume 1 of 3

Editors:

J. Vanschoren
S. Yeung

ISBN: 978-1-7138-7109-5

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571



Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2021) by individual authors and Neural Information Processing Systems Foundation Inc.
All rights reserved.

Printed by Curran Associates, Inc. (2024)

For permission requests, please contact Neural Information Processing Systems Foundation Inc.
at the address below.

Neural Information Processing Systems Foundation Inc.
1269 Law Street
San Diego, CA 92109
USA

Phone: (858) 453-4100
Fax: (858) 587-0417

<https://neurips.cc/>

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

TABLE OF CONTENTS

VOLUME 1

A BILINGUAL, OPENWORLD VIDEO TEXT DATASET AND END-TO-END VIDEO TEXT SPOTTER WITH TRANSFORMER	1
<i>Debing Zhang, Yuanqiang Cai, Sibow Wang, Sibow Wang, Jiahong Li, Zhuang Li, Yejun Tang, Hong Zhou</i>	
A CHANNEL CODING BENCHMARK FOR META-LEARNING	15
<i>Rui Li, Ondrej Bohdal, Rajesh K Mishra, Hyeji Kim, Da Li, Nicholas Lane, Timothy Hospedales</i>	
A DATASET FOR ANSWERING TIME-SENSITIVE QUESTIONS.....	27
<i>Wenhu Chen, Xinyi Wang, William Yang Wang, William Yang Wang</i>	
A LARGE-SCALE DATABASE FOR GRAPH REPRESENTATION LEARNING	40
<i>Scott Freitas, Yuxiao Dong, Joshua Neil, Duen Horng Chau</i>	
A PROCEDURAL WORLD GENERATION FRAMEWORK FOR SYSTEMATIC EVALUATION OF CONTINUAL LEARNING	53
<i>Timm Hess, Martin Mundt, Iuliia Pliushch, Visvanathan Ramesh</i>	
A REALISTIC APPROACH TO GENERATE MASKED FACES APPLIED ON TWO NOVEL MASKED FACE RECOGNITION DATA SETS	66
<i>Tudor-Alexandru Mare, Georgian Duta, Iuliana Georgescu, Adrian Sandru, Bogdan Alexe, Marius Popescu, Radu Tudor Ionescu</i>	
A SANDBOX FOR PREDICTION AND INTEGRATION OF DNA, RNA, AND PROTEINS IN SINGLE CELLS.....	83
<i>Malte Luecken, Daniel Burkhardt, Robrecht Cannoodt, Christopher Lance, Aditi Agrawal, Hananeh Aliee, Ann Chen, Louise Deconinck, Angela Detweiler, Alejandro Granados, Shelly Huynh, Laura Isacco, Yang Kim, Dominik Klein, Bony De Kumar, Sunil Kuppasani, Heiko Lickert, Aaron McGeever, Joaquin Melgarejo, Honey Mekonen, Maurizio Morri, Michaela Müller, Norma Neff, Sheryl Paul, Bastian Rieck, Kaylie Schneider, Scott Steelman, Michael Sterr, Daniel Treacy, Alexander Tong, Alexandra-Chloe Villani, Guilin Wang, Jia Yan, Ce Zhang, Angela Pisco, Smrita Krishnaswamy, Fabian Theis, Jonathan M Bloom</i>	
A SPOKEN LANGUAGE DATASET OF DESCRIPTIONS FOR SPEECH-BASED GROUNDED LANGUAGE LEARNING	96
<i>Gaoussou Kebe, Pdraig Higgins, Patrick Jenkins, Kasra Darvish, Rishabh Sachdeva, Ryan Barron, John Winder, Donald Engel, Edward Raff, Francis Ferraro, Cynthia Matuszek</i>	
A TOOLBOX FOR CONSTRUCTION AND ANALYSIS OF SPEECH DATASETS	110
<i>Evelina Bakhturina, Vitaly Lavrukhin, Boris Ginsburg</i>	
A UNIFIED FEW-SHOT CLASSIFICATION BENCHMARK TO COMPARE TRANSFER AND META LEARNING APPROACHES	123
<i>Vincent Dumoulin, Neil Houlsby, Utku Evci, Xiaohua Zhai, Ross Goroshin, Sylvain Gelly, Hugo Larochelle, Hugo Larochelle</i>	
ADDRESSING "DOCUMENTATION DEBT" IN MACHINE LEARNING: A RETROSPECTIVE DATASHEET FOR BOOKCORPUS.....	136
<i>John Bandy, Nicholas Vincent</i>	

ADVERSARIAL GLUE: A MULTI-TASK BENCHMARK FOR ROBUSTNESS EVALUATION OF LANGUAGE MODELS	149
<i>Boxin Wang, Chejian Xu, Shuohang Wang, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Awadallah, Bo Li</i>	
AI AND THE EVERYTHING IN THE WHOLE WIDE WORLD BENCHMARK	164
<i>Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, Amandalynne Paullada</i>	
ALCHEMY: A BENCHMARK AND ANALYSIS TOOLKIT FOR META-REINFORCEMENT LEARNING AGENTS	176
<i>Jane Wang, Michael King, Nicolas Porcel, Zeb Kurth-Nelson, Tina Zhu, Charles Deck, Peter Choy, Mary Cassin, Malcolm Reynolds, Francis Song, Gavin Buttimore, David Reichert, Neil Rabinowitz, Loic Matthey, Demis Hassabis, Alexander Lerchner, Matt Botvinick</i>	
AN EMPIRICAL INVESTIGATION OF REPRESENTATION LEARNING FOR IMITATION.....	190
<i>Cynthia Chen, Xin Chen, Sam Toyer, Cody Wild, Scott Emmons, Ian Fischer, Kuang-Huei Lee, Neel Alex, Steven Wang, Ping Luo, Stuart Russell, Pieter Abbeel, Rohin Shah</i>	
AN EMPIRICAL STUDY OF GRAPH CONTRASTIVE LEARNING	201
<i>Yanqiao Zhu, Yichen Xu, Qiang Liu, Shu Wu</i>	
AN EXTENSIBLE BENCHMARK SUITE FOR LEARNING TO SIMULATE PHYSICAL SYSTEMS.....	220
<i>Karl Otness, Arvi Gjoka, Joan Bruna, Daniele Panozzo, Benjamin Peherstorfer, Teseo Schneider, Denis Zorin</i>	
AN INFORMATION RETRIEVAL APPROACH TO BUILDING DATASETS FOR HATE SPEECH DETECTION.....	233
<i>Md Mustafizur Rahman, Dinesh Balakrishnan, Dhiraj Murthy, Mucahid Kutlu, Matt Lease</i>	
AP-10K: A BENCHMARK FOR ANIMAL POSE ESTIMATION IN THE WILD	245
<i>Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, Dacheng Tao</i>	
ARE WE LEARNING YET? A META REVIEW OF EVALUATION FAILURES ACROSS MACHINE LEARNING	275
<i>Thomas Liao, Rohan Taori, Deborah Raji, Ludwig Schmidt</i>	
ARGOVERSE 2: NEXT GENERATION DATASETS FOR SELF-DRIVING PERCEPTION AND FORECASTING	288
<i>Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, James Hays</i>	
ARKITSCENES: A DIVERSE REAL-WORLD DATASET FOR 3D INDOOR SCENE UNDERSTANDING USING MOBILE RGB-D DATA	303
<i>Afshin Dehghan, Gilad Baruch, Zhuoyuan Chen, Yuri Feigin, Peter Fu, Thomas Gebauer, Daniel Kurz, Tal Dimry, Brandon Joffe, Arik Schwartz, Elad Shulman</i>	
ARTSHEETS FOR ART DATASETS	318
<i>Ramya Srinivasan, Emily Denton, Jordan Famularo, Negar Rostamzadeh, Fernando Diaz, Beth Coleman</i>	
ATOM3D: TASKS ON MOLECULES IN THREE DIMENSIONS.....	331
<i>Raphael Townshend, Martin Vögele, Patricia Suriana, Alex Derry, Alexander Powers, Gianni Laloudakis, Sidhika Balachandar, Bowen Jing, Brandon Anderson, Stephan Eismann, Risi Kondor, Russ Altman, Ron Dror</i>	

AUTOMATIC CONSTRUCTION OF EVALUATION SUITES FOR NATURAL LANGUAGE GENERATION DATASETS	345
<i>Simon Mille, Kaustubh Dhole, Saad Mahamood, Laura Perez-Beltrachini, Varun Prashant Gangal, Mihir Kale, Emiel van Miltenburg, Sebastian Gehrmann</i>	
BEIR: A HETEROGENEOUS BENCHMARK FOR ZERO-SHOT EVALUATION OF INFORMATION RETRIEVAL MODELS	359
<i>Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, Iryna Gurevych</i>	
BENCHMARK FOR COMPOSITIONAL TEXT-TO-IMAGE SYNTHESIS.....	375
<i>Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, Anna Rohrbach</i>	
BENCHMARKING BAYESIAN DEEP LEARNING ON DIABETIC RETINOPATHY DETECTION TASKS	388
<i>Neil Band, Tim G. J. Rudner, Qixuan Feng, Angelos Filos, Zachary Nado, Mike Dusenberry, Ghassen Jerfel, Dustin Tran, Yarin Gal</i>	
BENCHMARKING BIAS MITIGATION ALGORITHMS IN REPRESENTATION LEARNING THROUGH FAIRNESS METRICS	403
<i>Charan Reddy, Deepak Sharma, Soroush Mehri, Adriana Romero Soriano, Samira Shabani, Sina Honari</i>	
BENCHMARKING DATA-DRIVEN SURROGATE SIMULATORS FOR ARTIFICIAL ELECTROMAGNETIC MATERIALS.....	437
<i>Yang Deng, Juncheng Dong, Simiao Ren, Omar Khatib, Mohammadreza Soltani, Vahid Tarokh, Willie Padilla, Jordan Malof</i>	
BENCHMARKING MULTI-AGENT DEEP REINFORCEMENT LEARNING ALGORITHMS IN COOPERATIVE TASKS.....	450
<i>Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, Stefano Albrecht</i>	
BENCHMARKING MULTIMODAL AUTOML FOR TABULAR DATA WITH TEXT FIELDS	463
<i>Xingjian Shi, Jonas Mueller, Nick Erickson, Nick Erickson, Mu Li, Alexander Smola, Alexander Smola</i>	
BENCHMARKING THE COMBINATORIAL GENERALIZABILITY OF COMPLEX QUERY ANSWERING ON KNOWLEDGE GRAPHS	477
<i>Zihao Wang, Hang Yin, Yangqiu Song</i>	
BENCHMARKING THE ROBUSTNESS OF SPATIAL-TEMPORAL MODELS AGAINST CORRUPTIONS	489
<i>Chenyu Yi, Siyuan Yang, Haoliang Li, Yap-peng Tan, Alex Kot</i>	
BENCHMARKS FOR CORRUPTION INVARIANT PERSON RE-IDENTIFICATION	501
<i>Minghui Chen, Zhiqiang Wang, Feng Zheng</i>	
BITOD: A BILINGUAL MULTI-DOMAIN DATASET FOR TASK-ORIENTED DIALOGUE MODELING.....	514
<i>Zhaojiang Lin, Andrea Madotto, Genta Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, Pascale N Fung</i>	
B-PREF: BENCHMARKING PREFERENCE-BASED REINFORCEMENT LEARNING	528
<i>Kimin Lee, Laura Smith, Anca Dragan, Pieter Abbeel</i>	
BRAX - A DIFFERENTIABLE PHYSICS ENGINE FOR LARGE SCALE RIGID BODY SIMULATION.....	543
<i>Daniel Freeman, Erik Frey, Anton Raichuk, Sertan Girgin, Igor Mordatch, Olivier Bachem</i>	
CARLA: A PYTHON LIBRARY TO BENCHMARK ALGORITHMIC RECOURSE AND COUNTERFACTUAL EXPLANATION ALGORITHMS	556
<i>Martin Pawelczyk, Sascha Bielawski, Johan Van den Heuvel, Tobias Richter, Gjergji Kasneci</i>	

CCNLAB: A BENCHMARKING FRAMEWORK FOR COMPUTATIONAL COGNITIVE NEUROSCIENCE.....	569
<i>Nikhil Bhattasali, Momchil Tomov, Samuel J Gershman</i>	
CHAOS AS AN INTERPRETABLE BENCHMARK FOR FORECASTING AND DATA-DRIVEN MODELLING	582
<i>William Gilpin</i>	
CHEST IMAGENOME DATASET FOR CLINICAL REASONING	598
<i>Joy T Wu, Nkechinyere Agu, Ismini Lourentzou, Ismini Lourentzou, Arjun Sharma, Joseph Alexander Paguio, Jasper Seth Yao, Edward C Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo Anthony Celi, Mehdi Moradi</i>	
CLEVRTEX: A TEXTURE-RICH BENCHMARK FOR UNSUPERVISED MULTI-OBJECT SEGMENTATION	612
<i>Laurynas Karazija, Iro Laina, Christian Rupprecht</i>	
CLIMART: A BENCHMARK DATASET FOR EMULATING ATMOSPHERIC RADIATIVE TRANSFER IN WEATHER AND CLIMATE MODELS.....	626
<i>Salva Rühling Cachay, Venkatesh Ramesh, Jason Cole, Howard Barker, David Rolnick</i>	
CODENET: A LARGE-SCALE AI FOR CODE DATASET FOR LEARNING A DIVERSITY OF CODING TASKS	638
<i>Ruchir Puri, David Kung, Geert Janssen, Wei Zhang, Giacomo Domeniconi, Vladimir Zolotov, Julian T Dolby, Jie Chen, Mihir Choudhury, Lindsey Decker, Veronika Thost, Veronika Thost, Luca Buratti, Saurabh Pujar, Shyam Ramji, Ulrich Finkler, Susan Malaika, Frederick Reiss</i>	
CODEXGLUE: A MACHINE LEARNING BENCHMARK DATASET FOR CODE UNDERSTANDING AND GENERATION.....	651
<i>Shuai Lu, Daya Guo, Shuo Ren, Junjie Huang, Alexey Svyatkovskiy, Ambrosio Blanco, Colin Clement, Dawn Drain, Daxin Jiang, Duyu Tang, Ge Li, Lidong Zhou, Linjun Shou, Long Zhou, Michele Tufano, Ming Gong, Ming Zhou, Nan Duan, Neel Sundaresan, Shao Kun Deng, Shengyu Fu, Shujie Liu</i>	
COMMONSENSEQA 2.0: EXPOSING THE LIMITS OF AI THROUGH GAMIFICATION	667
<i>Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, Jonathan Berant</i>	
CONSTRUCTING A VISUAL DATASET TO STUDY THE EFFECTS OF SPATIAL APARTHEID IN SOUTH AFRICA	681
<i>Raeeshe Sefala, Timnit Gebru, Nyalleng Moorosi, Luzango Mfupe, Richard Klein</i>	
CONTEMPORARY SYMBOLIC REGRESSION METHODS AND THEIR RELATIVE PERFORMANCE	695
<i>William La Cava, Patryk Orzechowski, Bogdan Burlacu, Fabricio de Franca, Marco Virgolin, Ying Jin, Michael Kommenda, Jason Moore</i>	
COVID-19 SOUNDS: A LARGE-SCALE AUDIO DATASET FOR DIGITAL RESPIRATORY SCREENING.....	711
<i>Tong Xia, Dimitris Spathis, Chloe Brown, J Ch, Andreas Grammenos, Jing Han, Apinan Hasthanasombat, Erika Bondareva, Ting Dang, Andres Floto, Pietro Cicuta, Cecilia Mascolo</i>	
CREAK: A DATASET FOR COMMONSENSE REASONING OVER ENTITY KNOWLEDGE	724
<i>Yasumasa Onoe, Michael Zhang, Eunsol Choi, Greg Durrett</i>	
CROPHARVEST: A GLOBAL DATASET FOR CROP-TYPE CLASSIFICATION	737
<i>Gabriel Tseng, Ivan Zvonkov, Catherine Nakalembe, Hannah Kerner</i>	

CROWDSPEECH AND VOX DIY: BENCHMARK DATASET FOR CROWDSOURCED AUDIO TRANSCRIPTION	751
<i>Nikita Pavlichenko, Ivan Stelmakh, Dmitry Ustalov</i>	
CSAW-M: AN ORDINAL CLASSIFICATION DATASET FOR BENCHMARKING MAMMOGRAPHIC MASKING OF CANCER	765
<i>Moein Sorkhei, Yue Liu, Hossein Azizpour, Edward Azavedo, Karin Dembrower, Dimitra Ntoula, Athanasios Zouzos, Fredrik Strand, Kevin Smith</i>	
CSFCUBE - A TEST COLLECTION OF COMPUTER SCIENCE RESEARCH ARTICLES FOR FACETED QUERY BY EXAMPLE	777
<i>Sheshera Mysore, Tim O’Gorman, Andrew McCallum, Hamed Zamani</i>	
CUAD: AN EXPERT-ANNOTATED NLP DATASET FOR LEGAL CONTRACT REVIEW	793
<i>Dan Hendrycks, Collin Burns, Anya Chen, Spencer Ball</i>	
DABS: A DOMAIN-AGNOSTIC BENCHMARK FOR SELF-SUPERVISED LEARNING	803
<i>Alex Tamkin, Vincent Liu, Rongfei Lu, Daniel Fein, Colin Schultz, Noah Goodman</i>	

VOLUME 2

DATASETS FOR ONLINE CONTROLLED EXPERIMENTS	825
<i>Chak Hin Bryan Liu, Angelo Cardoso, Paul Couturier, Emma McCoy</i>	
DEBAGREEMENT: A COMMENT-REPLY DATASET FOR (DIS)AGREEMENT DETECTION IN ONLINE DEBATES	842
<i>John Poug��-Biyong, Valentina Semenova, Alexandre Matton, Rachel Han, Aerin Kim, Renaud Lambiotte, Doyne Farmer</i>	
DENETHOR: THE DYNAMIC EARTHNET DATASET FOR HARMONIZED, INTER- OPERABLE, ANALYSIS-READY, DAILY CROP MONITORING FROM SPACE	854
<i>Lukas Kondmann, Aysim Toker, Marc Ru��wurm, Andr��s Camero, Devis Peressuti, Grega Milcinski, Pierre-Philippe Mathieu, Nicolas Longepe, Timothy Davis, Giovanni Marchisio, Laura Leal-Taix��, Xiaoxiang Zhu</i>	
DUE: END-TO-END DOCUMENT UNDERSTANDING BENCHMARK	867
<i>Lukasz Borchmann, Micha��l Pietruszka, Tomasz Stanislawek, Dawid Jurkiewicz, Micha��l Turski, Karolina Szyndler, Filip Grali��ski</i>	
DYNAMIC ENVIRONMENTS WITH DEFORMABLE OBJECTS	881
<i>Rika Antonova, peiyang shi, Hang Yin, Zehang Weng, Danica Kragic</i>	
EEGEYENET: A SIMULTANEOUS ELECTROENCEPHALOGRAPHY AND EYE-TRACKING DATASET AND BENCHMARK FOR EYE MOVEMENT PREDICTION	895
<i>Ard Kastrati, Martyna Plomecka, Damian Pascual Ortiz, Lukas Wolf, Victor Gillioz, Roger Wattenhofer, Nicolas Langer</i>	
EMPIRICAL STUDY OF OFF-POLICY POLICY EVALUATION FOR REINFORCEMENT LEARNING.....	910
<i>Cameron Voloshin, Hoang Le, Nan Jiang, Yisong Yue</i>	
EVALUATING BAYES ERROR ESTIMATORS ON REAL-WORLD DATASETS WITH FEEBEE	922
<i>Cedric Renggli, Luka Rimanic, Nora Hollenstein, Ce Zhang</i>	
EVENTNARRATIVE: A LARGE-SCALE EVENT-CENTRIC DATASET FOR KNOWLEDGE GRAPH-TO-TEXT GENERATION	933
<i>Anthony Colas, Ali Sadeghian, Yue Wang, Daisy Zhe Wang</i>	

FAKEAVCELEB: A NOVEL AUDIO-VIDEO MULTIMODAL DEEPPFAKE DATASET	946
<i>Hasam Khalid, Shahroz Tariq, Minha Kim, Simon Woo</i>	
FEVEROUS: FACT EXTRACTION AND VERIFICATION OVER UNSTRUCTURED AND STRUCTURED INFORMATION	960
<i>Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal</i>	
FEW-SHOT LEARNING EVALUATION IN NATURAL LANGUAGE UNDERSTANDING	974
<i>Subhabrata Mukherjee, Xiaodong Liu, Guoqing Zheng, Saghar Hosseini, Saghar Hosseini, Hao Cheng, Ge Yang, Christopher Meek, Ahmed Awadallah, Jianfeng Gao</i>	
FFA-IR: TOWARDS AN EXPLAINABLE AND RELIABLE MEDICAL REPORT GENERATION BENCHMARK	988
<i>Mingjie Li, Wenjia Cai, Rui Liu, Yuetian Weng, Xiaoyun Zhao, Cong Wang, Xin Chen, Zhong Liu, Caineng Pan, Mengke Li, yingfeng zheng, Yizhi Liu, Flora Salim, Karin Verspoor, Xiaodan Liang, Xiaojun Chang</i>	
FLIP: BENCHMARK TASKS IN FITNESS LANDSCAPE INFERENCE FOR PROTEINS	1001
<i>Christian Dallago, Jody Mou, Jody Mou, Kadina Johnston, Bruce Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, Kevin Yang</i>	
FS-MOL: A FEW-SHOT LEARNING DATASET OF MOLECULES	1015
<i>Megan Stanley, John Bronskill, Krzysztof Maziarz, Hubert Misztela, Jessica Lanini, Marwin Segler, Nadine Schneider, Marc Brockschmidt</i>	
GENERATING DATASETS OF 3D GARMENTS WITH SEWING PATTERNS	1028
<i>Maria Korosteleva, Sung-Hee Lee</i>	
GRAPH ROBUSTNESS BENCHMARK: BENCHMARKING THE ADVERSARIAL ROBUSTNESS OF GRAPH MACHINE LEARNING	1038
<i>Qinkai Zheng, Xu Zou, Yuxiao Dong, Yukuo Cen, Da Yin, Jiarong Xu, Jiarong Xu, Yang Yang, Jie Tang</i>	
GRAPHGT: MACHINE LEARNING DATASETS FOR GRAPH GENERATION AND TRANSFORMATION	1055
<i>Yuanqi Du, Shiyu Wang, Xiaojie Guo, Hengning Cao, Shujie Hu, Junji Jiang, Aishwarya Varala, Abhinav Angirekula, Liang Zhao</i>	
HABITAT-MATTERPORT 3D DATASET (HM3D): 1000 LARGE-SCALE 3D ENVIRONMENTS FOR EMBODIED AI	1069
<i>Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel Chang, Manolis Savva, Yili Zhao, Dhruv Batra</i>	
HARDWARE DESIGN AND ACCURATE SIMULATION OF STRUCTURED-LIGHT SCANNING FOR BENCHMARKING OF 3D RECONSTRUCTION ALGORITHMS	1081
<i>Sebastian Koch, Yurii Piadyk, Markus Worchel, Marc Alexa, Claudio Silva, Denis Zorin, Daniele Panozzo</i>	
HIRID-ICU-BENCHMARK --- A COMPREHENSIVE MACHINE LEARNING BENCHMARK ON HIGH-RESOLUTION ICU DATA	1093
<i>Hugo Yèche, Rita Kuznetsova, Marc Zimmermann, Matthias Hüser, Xinrui Lyu, Martin Faltsy, Gunnar Rätsch</i>	
HPO-B: A LARGE-SCALE REPRODUCIBLE BENCHMARK FOR BLACK-BOX HPO BASED ON OPENML	1105
<i>Sebastian Pineda Arango, Hadi Jomaa, Martin Wistuba, Josif Grabocka</i>	

HPOBENCH: A COLLECTION OF REPRODUCIBLE MULTI-FIDELITY BENCHMARK PROBLEMS FOR HPO	1117
<i>Katharina Eggensperger, Philipp Müller, Neeratoy Mallik, Matthias Feurer, Rene Sass, Aaron Klein, Noor Awad, Marius Lindauer, Frank Hutter</i>	
HUMBUGDB: A LARGE-SCALE ACOUSTIC MOSQUITO DATASET	1134
<i>Ivan Kiskin, Marianne Sinka, Adam Cobb, Waqas Rafique, Lawrence Wang, Davide Zilli, Benjamin Gutteridge, Rinita Dam, Theodoros Marinos, Yunpeng Li, Dickson Msaky, Emmanuel Kaindoa, Gerard Killeen, Eva Herreros-Moya, Kathy Willis, Stephen J Roberts</i>	
ICONQA: A NEW BENCHMARK FOR ABSTRACT DIAGRAM UNDERSTANDING AND VISUAL LANGUAGE REASONING	1147
<i>Pan Lu, Liang Qiu, Jiaqi Chen, Tanglin Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, Song-Chun Zhu</i>	
IMAGENET-21K PRETRAINING FOR THE MASSES	1161
<i>Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, Lihi Zelnik</i>	
INTELLIGENT SIGHT AND SOUND: A CHRONIC CANCER FACIAL PAIN DATASET	1173
<i>Catherine Ordun, Alexandra Cha, Edward Raff, Byron Gaskin, Alexander Hanson, Mason Rule, Sanjay Purushotham, James Gulley</i>	
ISAAC GYM: HIGH PERFORMANCE GPU BASED PHYSICS SIMULATION FOR ROBOT LEARNING.....	1186
<i>Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, Gavriel State</i>	
IT'S COMPASLICATED: THE MESSY RELATIONSHIP BETWEEN RAI DATASETS AND ALGORITHMIC FAIRNESS BENCHMARKS	1198
<i>Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, Suresh Venkatasubramanian</i>	
KESPEECH: AN OPEN SOURCE SPEECH DATASET OF MANDARIN AND ITS EIGHT SUBDIALECTS	1216
<i>Zhiyuan Tang, Dong Wang, Yanguang Xu, Jianwei Sun, Xiaoning Lei, Shuaijiang Zhao, Chengwen, Xingjun Tan, Chuandong Xie, Shuran Zhou, Rui Yan, Chenjia Lv, Yang Han, Wei Zou, Xiangang Li</i>	
KLUE: KOREAN LANGUAGE UNDERSTANDING EVALUATION	1228
<i>Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Lucy Park, Alice Oh, Jung-Woo Ha, Kyunghyun Cho, Kyunghyun Cho</i>	
LIRO: BENCHMARK AND LEADERBOARD FOR ROMANIAN LANGUAGE TASKS	1253
<i>Stefan Dumitrescu, Petru Rebeja, Beata Lorincz, Mihaela Gaman, Andrei Avram, Mihai Ilie, Andrei Pruteanu, Adriana Stan, Lorena Rosia, Cristina Iacobescu, Luciana Morogan, George Dima, Gabriel Marchidan, Traian Rebedea, Madalina Chitez, Dani Yogatama, Sebastian Ruder, Radu Tudor Ionescu, Razvan Pascanu, Viorica Patraucean</i>	
LOVEDA: A REMOTE SENSING LAND-COVER DATASET FOR DOMAIN ADAPTIVE SEMANTIC SEGMENTATION	1266
<i>Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, Yanfei Zhong</i>	
MANISKILL: GENERALIZABLE MANIPULATION SKILL BENCHMARK WITH LARGE-SCALE DEMONSTRATIONS	1278
<i>Tongzhou Mu, Zhan Ling, Fanbo Xiang, Derek Yang, Xuanlin Li, Xuanlin Li, Stone Tao, Zhiao Huang, Zhiwei Jia, Hao Su</i>	

MEASURING CODING CHALLENGE COMPETENCE WITH APPS	1293
<i>Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, Jacob Steinhardt</i>	
MEASURING MATHEMATICAL PROBLEM SOLVING WITH THE MATH DATASET	1304
<i>Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, Jacob Steinhardt</i>	
MIND DATASET FOR DIET PLANNING AND DIETARY HEALTHCARE WITH MACHINE LEARNING: DATASET CREATION USING COMBINATORIAL OPTIMIZATION AND CONTROLLABLE GENERATION WITH DOMAIN EXPERTS.....	1315
<i>Changhun Lee, Soohyeok Kim, Sehwa Jeong, Chiehyeon Lim, Jayun Kim, Yeji Kim, Minyoung Jung</i>	
MINIHACK THE PLANET: A SANDBOX FOR OPEN-ENDED REINFORCEMENT LEARNING RESEARCH.....	1328
<i>Mikayel Samvelyan, Robert Kirk, Vitaly Kurin, Jack Parker-Holder, Minqi Jiang, Eric Hambro, Fabio Petroni, Heinrich Kuttler, Edward Grefenstette, Tim Rocktäschel</i>	
MITIGATING DATASET HARMS REQUIRES STEWARDSHIP: LESSONS FROM 1000 PAPERS.....	1342
<i>Kenneth Peng, Arunesh Mathur, Arvind Narayanan</i>	
MLPERF TINY BENCHMARK.....	1358
<i>Colby Banbury, Vijay Janapa Reddi, Peter Torelli, Nat Jeffries, Csaba Kiraly, Jeremy Holleman, Pietro Montino, David Kanter, Pete Warden, Danilo Pau, Urmish Thakker, Antonio Torrini, Jay Cordaro, Giuseppe Di Guglielmo, Javier Duarte, Honson Tran, Nhan Tran, Niu Wenxu, Xu Xuesong</i>	
MODELING WORLDS IN TEXT.....	1373
<i>Prithviraj Ammanabrolu, Mark Riedl</i>	
RP-MOD & RP-CROWD: MODERATOR- AND CROWD-ANNOTATED GERMAN NEWS COMMENT DATASETS	1392
<i>Dennis Assenmacher, Marco Niemann, Kilian Müller, Moritz Seiler, Dennis Riehle, Heike Trautmann, Heike Trautmann</i>	
MONASH TIME SERIES FORECASTING ARCHIVE	1406
<i>Rakshitha W Godahewa, Christoph Bergmeir, Geoffrey Webb, Rob Hyndman, Pablo Montero-Manso</i>	
MQBENCH: TOWARDS REPRODUCIBLE AND DEPLOYABLE MODEL QUANTIZATION BENCHMARK	1432
<i>Yuhang Li, Mingzhu Shen, Jian Ma, Yan Ren, Mingxin Zhao, Qi Zhang, Ruihao Gong, Fengwei Yu, Junjie Yan</i>	
MULTIBENCH: MULTISCALE BENCHMARKS FOR MULTIMODAL REPRESENTATION LEARNING.....	1452
<i>Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie (Yufan) Chen, Peter Wu, Michelle A. Lee, Yuke Zhu, Ruslan Salakhutdinov, Louis-Philippe Morency</i>	
MULTILINGUAL SPOKEN WORDS CORPUS	1469
<i>Mark Mazumder, Sharad Chitlangia, Colby Banbury, Yiping Kang, Juan Ciro, Keith Achorn, Daniel Galvez, Mark Sabini, Peter Mattson, David Kanter, Greg Damos, Pete Warden, Josh Meyer, Vijay Janapa Reddi</i>	
NATIVE CHINESE READER: A DATASET TOWARDS NATIVE-LEVEL CHINESE MACHINE READING COMPREHENSION	1482
<i>Shusheng Xu, Yichen Liu, Xiaoyu Yi, Siyuan Zhou, Huizi Li, Yi Wu</i>	

NATURALPROOFS: MATHEMATICAL THEOREM PROVING IN NATURAL LANGUAGE	1496
<i>Sean Welleck, Jiacheng Liu, Ronan Le Bras, Hanna Hajishirzi, Yejin Choi, Kyunghyun Cho, Kyunghyun Cho</i>	
NATURE: NATURAL AUXILIARY TEXT UTTERANCES FOR REALISTIC SPOKEN LANGUAGE EVALUATION	1509
<i>David Alfonso-Hermelo, Ahmad Rashid, Abbas Ghaddar, Philippe Langlais, Mehdi Rezagholizadeh</i>	
NEURAL LATENTS BENCHMARK '21: EVALUATING LATENT VARIABLE MODELS OF NEURAL POPULATION ACTIVITY	1526
<i>Felix Pei, Joel Ye, David Zoltowski, David Zoltowski, Anqi Wu, Raaed Chowdhury, Hansem Sohn, Joseph O'Doherty, Krishna V Shenoy, Matthew Kaufman, Mark Churchland, Mehrdad Jazayeri, Lee Miller, Jonathan Pillow, Il Memming Park, Eva Dyer, Chethan Pandarinath</i>	
OCCLUDED VIDEO INSTANCE SEGMENTATION: DATASET AND ICCV 2021 CHALLENGE	1537
<i>Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip Torr, Song Bai</i>	
OGB-LSC: A LARGE-SCALE CHALLENGE FOR MACHINE LEARNING ON GRAPHS.....	1552
<i>Weihua Hu, Matthias Fey, Hongyu Ren, Maho Nakata, Yuxiao Dong, Jure Leskovec</i>	
OMNIPRINT: A CONFIGURABLE PRINTED CHARACTER SYNTHESIZER	1567
<i>Haozhe Sun, Wei-Wei Tu, Isabelle Guyon</i>	
ONE MILLION SCENES FOR AUTONOMOUS DRIVING: ONCE DATASET.....	1580
<i>Jiageng Mao, Niu Minzhe, ChenHan Jiang, hanxue liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Jie Yu, Chunjing XU, Hang Xu</i>	
OPEN BANDIT DATASET AND PIPELINE: TOWARDS REALISTIC AND REPRODUCIBLE OFF-POLICY EVALUATION	1594
<i>Yuta Saito, Shunsuke Aihara, Megumi Matsutani, Yusuke Narita</i>	
OPENML BENCHMARKING SUITES.....	1608
<i>Bernd Bischl, Bernd Bischl, Giuseppe Casalicchio, Matthias Feurer, Pieter Gijsbers, Frank Hutter, Michel Lang, Rafael Gomes Mantovani, Jan van Rijn, Joaquin Vanschoren</i>	
PASS: AN IMAGENET REPLACEMENT FOR SELF-SUPERVISED PRETRAINING WITHOUT HUMANS	1634
<i>Yuki Asano, Christian Rupprecht, Andrew Zisserman, Andrea Vedaldi</i>	
PERSONALIZED BENCHMARKING WITH THE LUDWIG BENCHMARKING TOOLKIT	1648
<i>Avanika Narayan, Piero Molino, Karan Goel, Willie Neiswanger, Christopher Ré</i>	

VOLUME 3

PERVASIVE LABEL ERRORS IN TEST SETS DESTABILIZE MACHINE LEARNING BENCHMARKS	1661
<i>Curtis Northcutt, Anish Athalye, Jonas Mueller</i>	
PHYSION: EVALUATING PHYSICAL PREDICTION FROM VISION IN HUMANS AND MACHINES	1676
<i>Daniel Bear, Elias Wang, Damian Mrowca, Felix Binder, Hsiao-Yu Tung, Pramod RT, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, Fei-Fei Li, Nancy Kanwisher, Josh Tenenbaum, Dan Yamins, Judith Fan</i>	

PL@NTNET-300K: A PLANT IMAGE DATASET WITH HIGH LABEL AMBIGUITY AND A LONG-TAILED DISTRIBUTION.....	1689
<i>Camille Garcin, alexis joly, Pierre Bonnet, Antoine Affouard, Jean-Christophe Lombardo, Mathias Chouet, Maximilien Servajean, Titouan Lorieul, Joseph Salmon</i>	
PROCAT: PRODUCT CATALOGUE DATASET FOR IMPLICIT CLUSTERING, PERMUTATION LEARNING AND STRUCTURE PREDICTION.....	1702
<i>Mateusz Jurewicz, Leon Derczynski</i>	
PROGRAMMING PUZZLES.....	1717
<i>Tal Schuster, Ashwin Kalyan, Alex Polozov, Adam Kalai</i>	
Q-PAIN: A QUESTION ANSWERING DATASET TO MEASURE SOCIAL BIAS IN PAIN MANAGEMENT	1729
<i>Cécile Logé, Emily Ross, David Dadey, Saahil Jain, Adriel Saporta, Andrew Ng, Pranav Rajpurkar, Pranav Rajpurkar</i>	
RADGRAPH: EXTRACTING CLINICAL ENTITIES AND RELATIONS FROM RADIOLOGY REPORTS.....	1741
<i>Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew Lungren, Andrew Ng, Curtis Langlotz, Pranav Rajpurkar, Pranav Rajpurkar</i>	
RAFT: A REAL-WORLD FEW-SHOT TEXT CLASSIFICATION BENCHMARK	1753
<i>Neel Alex, Eli Lifland, Lewis Tunstall, Abhishek Thakur, Pegah Maham, C. Riedel, Emmie Hine, Carolyn Ashurst, Paul Sedille, Alexis Carlier, Michael Noetel, Andreas Stuhlmüller</i>	
RB2: ROBOTIC MANIPULATION BENCHMARKING WITH A TWIST	1766
<i>Sudeep Dasari, Jianren Wang, Joyce Hong, Shikhar Bahl, Yixin Lin, Austin Wang, Abitha Thankaraj, Karanbir Chahal, Berk Calli, Saurabh Gupta, David Held, Lerrel Pinto, Deepak Pathak, Vikash Kumar, Abhinav Gupta</i>	
REALLY DOING GREAT AT ESTIMATING CATE? A CRITICAL LOOK AT ML BENCHMARKING PRACTICES IN TREATMENT EFFECT ESTIMATION	1780
<i>Alicia Curth, David Svensson, Jim Weatherall, Mihaela van der Schaar</i>	
REASCAN: COMPOSITIONAL REASONING IN LANGUAGE GROUNDING.....	1793
<i>Zhengxuan Wu, Elisa Kreiss, Desmond Ong, Christopher Potts</i>	
REDCAPS: WEB-CURATED IMAGE-TEXT DATA CREATED BY THE PEOPLE, FOR THE PEOPLE	1807
<i>Karan Desai, Gaurav Kaul, Zubin Aysola, Justin Johnson</i>	
REDUCED, REUSED AND RECYCLED: THE LIFE OF A DATASET IN MACHINE LEARNING RESEARCH.....	1820
<i>Bernard Koch, Emily Denton, Alex Hanna, Jacob G Foster</i>	
REINFORCEMENT LEARNING BENCHMARKS FOR TRAFFIC SIGNAL CONTROL.....	1831
<i>James Ault, Guni Sharon</i>	
RELATIONAL PATTERN BENCHMARKING ON THE KNOWLEDGE GRAPH LINK PREDICTION TASK	1843
<i>Afshin Sadeghi, Hirra Malik, Diego Collarana, Jens Lehmann</i>	
RELLISUR: A REAL LOW-LIGHT IMAGE SUPER-RESOLUTION DATASET	1855
<i>Andreas Aakerberg, Kamal Nasrollahi, Thomas Moeslund</i>	
REVISITING TIME SERIES OUTLIER DETECTION: DEFINITIONS AND BENCHMARKS	1868
<i>Kwei-Herng Lai, Daochen Zha, Junjie Xu, Yue Zhao, Guanchu Wang, Xia Hu</i>	

ROBUSTBENCH: A STANDARDIZED ADVERSARIAL ROBUSTNESS BENCHMARK.....	1885
<i>Francesco Croce, Maksym Andriushchenko, Vikash Sehwal, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, Matthias Hein</i>	
SCIGEN: A DATASET FOR REASONING-AWARE TEXT GENERATION FROM SCIENTIFIC TABLES	1898
<i>Nafise Moosavi, Andreas Rücklé, Dan Roth, Iryna Gurevych</i>	
SEASONS IN DRIFT: A LONG TERM THERMAL IMAGING DATASET FOR STUDYING CONCEPT DRIFT	1913
<i>Ivan Nikolov, Mark Philip Philipsen, Jinsong Liu, Jacob Dueholm, Anders Johansen, Kamal Nasrollahi, Thomas Moeslund</i>	
SEGMENTMEIFYOUCAN: A BENCHMARK FOR ANOMALY SEGMENTATION	1926
<i>Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Pascal Fua, Mathieu Salzmann, Matthias Rottmann</i>	
SHIFTS: A DATASET OF REAL DISTRIBUTIONAL SHIFT ACROSS MULTIPLE LARGE-SCALE TASKS	1941
<i>Andrey Malinin, Neil Band, Yarin Gal, Mark Gales, Alexander Ganshin, German Chesnokov, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, Vyas Raina, Denis Roginskiy, Mariya Shmatova, Panagiotis Tigas, Boris Yangel</i>	
SKM-TEA: A DATASET FOR ACCELERATED MRI RECONSTRUCTION WITH DENSE IMAGE LABELS FOR QUANTITATIVE CLINICAL EVALUATION	1955
<i>Arjun Desai, Andrew Schmidt, Elka Rubin, Christopher Sandino, Marianne Black, Valentina Mazzoli, Kathryn Stevens, Robert Boutin, Christopher Ré, Garry Gold, Brian Hargreaves, Akshay Chaudhari</i>	
SODA10M: A LARGE-SCALE 2D SELF/SEMI-SUPERVISED OBJECT DETECTION DATASET FOR AUTONOMOUS DRIVING	1968
<i>Jianhua Han, Xiwen Liang, Hang Xu, Kai Chen, Lanqing Hong, Jiageng Mao, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Xiaodan Liang, Chunjing Xu</i>	
STAR: A BENCHMARK FOR SITUATED REASONING IN REAL-WORLD VIDEOS.....	1981
<i>Bo Wu, Shoubin Yu, Zhenfang Chen, Josh Tenenbaum, Chuang Gan</i>	
STEP: SEGMENTING AND TRACKING EVERY PIXEL.....	1994
<i>Mark Weber, Jun Xie, Maxwell Collins, Yukun Zhu, Paul Voigtlaender, Hartwig Adam, Bradley Green, Andreas Geiger, Bastian Leibe, Daniel Cremers, Aljoša Ošep, Laura Leal-Taixé, Liang-Chieh Chen</i>	
SUSTAINBENCH: BENCHMARKS FOR MONITORING THE SUSTAINABLE DEVELOPMENT GOALS WITH MACHINE LEARNING	2011
<i>Christopher Yeh, Chenlin Meng, Sherrie Wang, Anne Driscoll, Erik Rozi, Patrick Liu, Jihyeon Lee, Marshall Burke, David Lobell, Stefano Ermon</i>	
SYNTHBIO: A CASE STUDY IN FASTER CURATION OF TEXT DATASETS	2035
<i>Ann Yuan, Daphne Ippolito, Vitaly Nikolaev, Chris Callison-Burch, Andy Coenen, Sebastian Gehrmann</i>	
SYNTHETIC BENCHMARKS FOR SCIENTIFIC RESEARCH IN EXPLAINABLE MACHINE LEARNING.....	2049
<i>Yang Liu, Sujay Khandagale, Sujay Khandagale, Colin White, Willie Neiswanger</i>	
SYSTEMATIC EVALUATION OF CAUSAL DISCOVERY IN VISUAL MODEL BASED REINFORCEMENT LEARNING	2062
<i>Nan Rosemary Ke, Aniket Didolkar, Sarthak Mittal, Anirudh Goyal, Guillaume Lajoie, Stefan Bauer, Danilo Jimenez Rezende, Michael Mozer, Yoshua Bengio, Chris Pal</i>	

TASK AGNOSTIC AND TASK SPECIFIC SELF-SUPERVISED LEARNING FROM SPEECH WITH LEBENCHMARK	2078
<i>Solène Evain, Ha Nguyen, Hang Le, Marceley Zanon Boito, Salima Mdhaffar, Sina Alisamir, Ziyi Tong, Natalia Tomashenko, Marco Dinarelli, Titouan Parcollet, Alexandre Allauzen, Yannick Estève, Benjamin Lecouteux, François Portet, Solange Rossato, Fabien Ringeval, Didier Schwab, Laurent Besacier</i>	
TEACH ME TO EXPLAIN: A REVIEW OF DATASETS FOR EXPLAINABLE NATURAL LANGUAGE PROCESSING.....	2101
<i>Sarah Wiegreffe, Ana Marasovic</i>	
TENSET: A LARGE-SCALE PROGRAM PERFORMANCE DATASET FOR LEARNED TENSOR COMPILERS	2115
<i>Lianmin Zheng, Ruochen Liu, Junru Shao, Tianqi Chen, Joseph Gonzalez, Ion Stoica, Ameer Haj-Ali</i>	
THE CLEAR BENCHMARK: CONTINUAL LEARNING ON REAL-WORLD IMAGERY	2130
<i>Zhiqiu Lin, Jia Shi, Deepak Pathak, Deva Ramanan</i>	
THE CPD DATA SET: PERSONNEL, USE OF FORCE, AND COMPLAINTS IN THE CHICAGO POLICE DEPARTMENT.....	2143
<i>Thibaut Horel, Lorenzo Masoero, Raj Agrawal, Daria Roithmayr, Trevor Campbell</i>	
THE MEDKIT-LEARN(ING) ENVIRONMENT: MEDICAL DECISION MODELLING THROUGH SIMULATION	2158
<i>Alex Chan, Ioana Bica, Alihan Hüyük, Daniel Jarrett, Mihaela van der Schaar</i>	
THE MET DATASET: INSTANCE-LEVEL RECOGNITION FOR ARTWORKS	2170
<i>Nikolaos-Antonios Ypsilantis, Noa Garcia, Guangxing Han, Sarah Ibrahimi, Nanne van Noord, Giorgos Tolias</i>	
THE MULTI-AGENT BEHAVIOR DATASET: MOUSE DYADIC SOCIAL INTERACTIONS	2182
<i>Jennifer J Sun, Tomomi Karigo, Dipam Chakraborty, Sharada Mohanty, Benjamin Wild, Quan Sun, Chen Chen, David Anderson, Pietro Perona, Yisong Yue, Ann Kennedy</i>	
THE NEURAL MMO PLATFORM FOR MASSIVELY MULTIAGENT RESEARCH	2197
<i>Joseph Suarez, Yilun Du, Clare Zhu, Igor Mordatch, Phillip Isola</i>	
THE PAIR-R24M DATASET FOR MULTI-ANIMAL 3D POSE ESTIMATION	2211
<i>Jesse Marshall, Ugne Klibaite, Amanda Gellis, Diego Aldarondo, Bence Olveczky, Timothy W Dunn</i>	
THE PEOPLE’S SPEECH: A LARGE-SCALE DIVERSE ENGLISH SPEECH RECOGNITION DATASET FOR COMMERCIAL USAGE.....	2224
<i>Daniel Galvez, Greg Diamos, Juan Torres, Keith Achorn, Juan Cerón, Anjali Gopi, David Kanter, Max Lam, Mark Mazumder, Vijay Janapa Reddi</i>	
THE TUFTS FNIRS MENTAL WORKLOAD DATASET & BENCHMARK FOR BRAIN-COMPUTER INTERFACES THAT GENERALIZE.....	2236
<i>Zhe Huang, Liang Wang, Giles Blaney, Christopher Slaughter, Devon McKeon, Ziyu Zhou, Robert Jacob, Robert Jacob, Michael Hughes</i>	
THERAPEUTICS DATA COMMONS: MACHINE LEARNING DATASETS AND TASKS FOR DRUG DISCOVERY AND DEVELOPMENT.....	2258
<i>Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor Coley, Cao Xiao, Jimeng Sun, Marinka Zitnik</i>	

THREEDWORLD: A PLATFORM FOR INTERACTIVE MULTI-MODAL PHYSICAL SIMULATION.....	2281
<i>Chuang Gan, Jeremy Schwartz, Seth Alter, Damian Mrowca, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaladar, Nick Haber, Megumi Sano, Kuno Kim, Elias Wang, Michael Lingelbach, Aidan Curtis, Kevin Feigelis, Daniel Bear, Dan Gutfreund, David Cox, Antonio Torralba, James J DiCarlo, Josh Tenenbaum, Josh McDermott, Dan Yamins</i>	
TIMERS AND SUCH: A PRACTICAL BENCHMARK FOR SPOKEN LANGUAGE UNDERSTANDING WITH NUMBERS.....	2294
<i>Loren Lugosch, Piyush Papreja, Mirco Ravanelli, Abdelwahab HEBA, Titouan Parcollet, Titouan Parcollet</i>	
TOWARDS A ROBUST EXPERIMENTAL FRAMEWORK AND BENCHMARK FOR LIFELONG LANGUAGE LEARNING	2305
<i>Aman Hussain, Nithin Holla, Pushkar Mishra, Helen Yannakoudakis, Ekaterina Shutova</i>	
TRUST, BUT VERIFY: CROSS-MODALITY FUSION FOR HD MAP CHANGE DETECTION.....	2321
<i>John Lambert, James Hays</i>	
URLB: UNSUPERVISED REINFORCEMENT LEARNING BENCHMARK.....	2335
<i>Misha Laskin, Denis Yarats, Hao Liu, Kimin Lee, Albert Zhan, Kevin Lu, Catherine Cang, Lerrel Pinto, Pieter Abbeel</i>	
VALUE: A MULTI-TASK BENCHMARK FOR VIDEO-AND-LANGUAGE UNDERSTANDING EVALUATION	2349
<i>Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Wang, William Yang Wang, William Yang Wang, Tamara L Berg, Mohit Bansal, Jingjing Liu, Lijuan Wang, Zicheng Liu</i>	
VARIANCE-AWARE MACHINE TRANSLATION TEST SETS	2362
<i>Runzhe Zhan, Xuebo Liu, Derek Wong, Lidia Chao</i>	
VFP290K: A LARGE-SCALE BENCHMARK DATASET FOR VISION-BASED FALLEN PERSON DETECTION	2376
<i>Jaegu An, Jeongho Kim, Hanbeen Lee, Jinbeom Kim, Junhyung Kang, Minha Kim, Saebyeol Shin, Minha Kim, Donghee Hong, Simon Woo</i>	
WAVEFAKE: A DATA SET TO FACILITATE AUDIO DEEPPAKE DETECTION	2389
<i>Joel Frank, Lea Schönherr</i>	
WHAT AILS ONE-SHOT IMAGE SEGMENTATION: A DATA PERSPECTIVE	2406
<i>Mayur Hemani, Abhinav Patel, Tejas Shimpi, Anirudha Ramesh, Balaji Krishnamurthy</i>	
WHAT WOULD JIMINY CRICKET DO? TOWARDS AGENTS THAT BEHAVE MORALLY	2419
<i>Dan Hendrycks, Mantas Mazeika, Andy Zou, Sahil Patel, Christine Zhu, Jesus Navarro, Dawn Song, Bo Li, Jacob Steinhardt</i>	
WHICH PRIORS MATTER? BENCHMARKING MODELS FOR LEARNING LATENT DYNAMICS.....	2431
<i>Aleksandar Botev, Andrew Jaegle, Peter Wirnsberger, Daniel Hennes, Irina Higgins</i>	
WHOLE BRAIN VESSEL GRAPHS: A DATASET AND BENCHMARK FOR GRAPH LEARNING AND NEUROSCIENCE	2445
<i>Johannes C. Paetzold, Julian McGinnis, Suprosanna Shit, Ivan Ezhov, Paul Büschl, Chinmay Prabhakar, Anjany Sekuboyina, Mihail Todorov, Georgios Kaissis, Ali Ertürk, Stephan Günemann, Bjoern Menze</i>	

WIKICHURCHES: A FINE-GRAINED DATASET OF ARCHITECTURAL STYLES WITH REAL-WORLD CHALLENGES.....	2458
<i>Björn Barz, Joachim Denzler</i>	
WILDFIREDB: AN OPEN-SOURCE DATASET CONNECTING WILDFIRE OCCURRENCE WITH RELEVANT DETERMINANTS	2471
<i>Samridhi Singla, Ayan Mukhopadhyay, Michael Wilbur, Tina Diao, Vinayak Gajjewar, Ahmed Eldawy, Mykel J Kochenderfer, Ross Shachter, Abhishek Dubey</i>	
WRENCH: A COMPREHENSIVE BENCHMARK FOR WEAK SUPERVISION	2479
<i>Jieyu Zhang, Yue Yu, NameError, Yujing Wang, Yaming Yang, Mao Yang, Alexander Ratner</i>	

Author Index