OMB No. 0925-0001 and 0925-0002 (Rev. 11/16 Approved Through 10/31/2018)

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.  
Follow this format for each person. **DO NOT EXCEED FIVE PAGES.**

NAME: William La Cava

eRA COMMONS USER NAME (credential, e.g., agency login): LACAVA

POSITION TITLE: Postdoctoral Researcher

EDUCATION/TRAINING

| INSTITUTION AND LOCATION | DEGREE  (if applicable) | Completion Date  MM/YYYY | FIELD OF STUDY |
| --- | --- | --- | --- |
| Cornell University (Ithaca, NY) | B.S. | 05/2009 | Mechanical & Aerospace Engineering |
| Cornell University (Ithaca, NY) | M.Eng. | 05/2010 | Mechanical & Aerospace Engineering |
| University of Massachusetts Amherst (Amherst, MA) | Ph.D. | 09/2016 | Mechanical & Industrial Engineering |
| University of Pennsylvania (Philadelphia, PA) | Postdoc |  | Biomedical Informatics |

**A. Personal Statement**

I am interested in developing methods for modeling complex systems that allow experts to better understand and control the processes they study. My general approach to this task is to focus on methods with high capacity for modeling non-linearity with as little structural complexity as possible. For most of my research career, I have applied these ideas to modeling dynamical systems. I have developed both stochastic and gradient-based methods of constructing and/or refining process models, with particular research focus on optimizing the structure of such models for conciseness. I have developed interpetable, data-driven models in several application areas, including wind energy, fluid dynamics and control. My publication record demonstrates my ability to conduct independent, high quality machine learning research (30 peer reviewed publications, 20 as first author), which led to my recruitment by Dr. Moore. During the my postdoctoral training with Dr. Moore, I have worked to develop techniques for the task of feature construction / representation learning by integrating model structure search with standard machine learning tools. In addition to methods development, I have been successful in producing three publications in the biomedical informatics field: two that benchmark machine learning methods and offer data-driven recommendations regarding their application to biomedical data, and one that reviews feature selection methods. I have produced 13 manuscripts in my first two years in the lab, 8 as first author. My long-term career goal is to secure a tenure-track faculty position and establish an interdiscplinary research lab at the intersection of machine learning and biomedical informatics. This proposal extends my preliminary work using concepts from multi-objective optimization and deep learning and applies these methods to the prediction of patient outcomes using electronic health record data. Electronic health records provide a wealth of information with high potential to infer drivers of patient outcomes. In particular, I plan to focus on predicting the risk of heart disease and heart failure, as well as risk of hospital readmissions. The proposed research will result in methods that optimize patient outcome models for accuracy as well as conciseness, and includes consultation plans with medical experts in order to assist the translation of these models into impactful changes in patient care. As part of my transition to independence, my research is a departure from Dr. Moore's research, which focuses primarily on identifying gene-gene and gene-environment interactions in human genomics. This proposal will provide me with expertise in mining electronic health records.

**B. Positions and Honors**

**Positions and Employment**

**2019 - Consultant, Division of Clinical Geriatrics and Gerontology, National Institute on Aging**

**2016 - Postdoctoral Fellow, Institute for Biomedical Informatics, University of Pennsylvania**

**2012 – 2016 PhD Student, University of Massachusetts Amherst**

2015 Visiting Researcher, Biosystems and Integrative Sciences Insitute (BioISI), University of Lisbon

2010 – 2012 Research Scientist, National Renewable Energy Laboratory (NREL)

2008 – 2010 Lead Engineer of Mechanical Power Systems, Cornell 100+ MPG Team

2007 – 2008 Independent Research, Cornell Computational Synthesis Laboratory

Other Experience and Professional Memberships

**2018 Organizer, GECCO Workshop on New Standards for Benchmarking**

2017 Guest Lecturer, machine learning unit of *Data Science For Biomedical Informatics*, University of Pennsylvania

**2017 Organizer, GECCO Workshop on Benchmarking Methods**

**2014 – 2016 Guest Lecturer and TA, *System Dynamics*; *Control Systems Laboratory*; *Offshore Wind Energy*; University of Massachusetts Amherst**

**2014- Member, Association of Computing Machinery (ACM)**

**2012 - 2014 American Institute of Aeronautics and Astronautics (AIAA)**

2011, 2012 Organizer, Gearbox Reliability Collaborative Workshop, National Renewable Energy Laboratory

2010- Member, American Society of Mechanical Engineers (ASME)

**Honors**

**2018 Best Paper Nomination, Genetic and Evolutionary Computation Conference (GECCO)**

**2017 Best Paper Nomination, European Conference on Genetic Programming (EuroGP)**

2016 Postdoctoral Fellowship, Warren Center for Network and Data Sciences

2015 Best Paper Nomination, Genetic and Evolutionary Computation Conference (GECCO)

2012 NSF Fellowship, IGERT: Offshore Wind Energy Program

2011 First Place, Cornell 100+ MPG Team, Green Grand Prix Competition

2010 Second stage selection, Cornell 100+ MPG Team, Automotive X-Prize

**C. Contributions to Science**

1. **A fundamental problem in science is the formulation of a model of a system when no *a priori* knowledge of the process is available. To this end, symbolic regression methods were developed to search for a model's structure, as well as its parameters, using a stochastic optimization technique known as genetic programming. However, symbolic regression exhibits a well-known phenomenon known as bloat, in which it produces overly complex solutions. We developed a method known as *epigenetic local search* [1a] that specifically optimizes model structures for conciseness during optimization via a novel developmental encoding.** We found that this method produced concise solutions that generalized better to test data. We applied it to the modeling of wind turbine dynamics [1b], as well as fluid dynamics and other non-linear dynamic systems [1c] and showed it could find true underlying physical principles. In comparison to neural network and auto-regressive models, our method produced much simpler and often more accurate models. I served as the primary contributor to all parts of this project.
   1. **La Cava, W.**, Helmuth, T., Spector, L., & Danai, K. (2015). Genetic Programming with Epigenetic Local Search. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)* (pp. 1055–1062). ACM Press. https://doi.org/10.1145/2739480.2754763
   2. **La Cava, W.**, Danai, K., & Spector, L. (2016). Inference of compact nonlinear dynamic models by epigenetic local search. *Engineering Applications of Artificial Intelligence*, *55*, 292–306. https://doi.org/10.1016/j.engappai.2016.07.004
   3. **La Cava, W.**, Danai, K., Spector, L., Fleming, P., Wright, A., & Lackner, M. (2016). Automatic identification of wind turbine models using evolutionary multiobjective optimization. *Renewable Energy*, *87*, *Part 2*, 892–902. https://doi.org/10.1016/j.renene.2015.09.068
2. A model’s quality depends completely on the data representation used to train it. For this reason, methods that can transform input data such that it better suits a given machine learning method can improve that method's predictive capacity. We showed that symbolic regression approaches can be competitive in the task of learning better data representations for standard ML tools [2a, 2b, 2c, 2d]. These approaches have nice properties, including 1) the ability to represent arbitrary nonlinear relations in the data, 2) independent scaling from the number of features in the raw data, 3) the ability to produce readable transformations. On a set of 20 classification problems, an ensemble technique [2b] outperformed 7 state-of-the-art ML methods trained on the raw data [2b]. Our most recent work [2d] produce state-of-the-art results across 100 real-world regression problems. I served as the primary contributor to all parts of this project. .
   1. **La Cava, W.**, & Moore, J. (2017). A General Feature Engineering Wrapper for Machine Learning Using epsilon-Lexicase Survival. In *Genetic Programming* (pp. 80–95). Springer, Cham. https://doi.org/10.1007/978-3-319-55696-3\_6
   2. **La Cava, W.**, & Moore, J. H. (2017). Ensemble representation learning: an analysis of fitness and survival for wrapper-based genetic programming methods. In *GECCO ’17: Proceedings of the 2017 Genetic and Evolutionary Computation Conference* (pp. 961–968). Berlin, Germany: ACM. https://doi.org/10.1145/3071178.3071215
   3. **La Cava, W.**, Silva, S., Vanneschi, L., Spector, L., & Moore, J. (2017). Genetic Programming Representations for Multi-dimensional Feature Learning in Biomedical Classification. In *Applications of Evolutionary Computation* (pp. 158–173). Springer, Cham. <https://doi.org/10.1007/978-3-319-55849-3_11>
   4. **La Cava, W.**, Singh, T.R., Taggart, J., Suri, S., & Moore, J.H. (2019). Learning Concise Representations for Regression by Evolving Networks of Trees. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1807.00981.
3. The biomedical informatics field has shown increasing interest in the development and use of machine learning methods for improving predictions of outcomes of interest. The main issue facing methods development is a lack of adequate benchmarking standards for comparing new methods to those available in literature. To tackle this issue, we created the Penn Machine Learning Benchmark suite (PMLB) [3a], a cleaned, standardized, and fetchable archive of hundreds of open-source benchmark datasets collected from around the web, including informatics applications from fundamental biology, genetics, and clinical decision making. We provided a comprehensive analysis of this suite using 14 open-source machine learning methods and rigorous statistical analysis. We performed bi-clustering to extract groups of datasets for which certain machine learning methods outperform others. We quantified the effect of hyperparameter tuning and model selection strategies. Finally, we provided a short list of recommendation to biomedical researchers as a starting point for modeling their data [3b]. The results contribute a much needed advance to the practice of machine learning development as well as use in informatics. More recently, we conducted a comprehensive comparison of state-of-the-art symbolic regression methods in comparison to established ML methods [3c]. In addition to this work, I have helped to review feature selection methods that are sensitive to interactions for reduced-order modeling of high-dimensional biomedical data [3d]. I was second author establishing the benchmark data set, and I shared first authorship with our paper in PSB. I designed the study of symbolic regression methods and analyzed the results.
   1. Olson, R. S., **La Cava, W.**, Orzechowski, P., Urbanowicz, R. J., & Moore, J. H. (2017). PMLB: A Large Benchmark Suite for Machine Learning Evaluation and Comparison. *BioData Mining*. Retrieved from https://arxiv.org/abs/1703.00512.
   2. Olson, R. S.\*, **La Cava, W.\***, Mustahsan, Z., Varik, A., & Moore, J. H. (2017). Data-driven Advice for Applying Machine Learning to Bioinformatics Problems. *Pacific Symposium on Biocomputing (PSB)*. Retrieved from http://arxiv.org/abs/1708.05070. \*contributed equally.
   3. Orzechowski, P., **La Cava, W.**, & Moore, J. H. (2018). Where are we now? A large benchmark study of recent symbolic regression methods. *GECCO ‘18: Proceedings of the 2018 Genetic and Evolutionary Computation Conference. ArXiv:1804.09331 [Cs]*. <https://doi.org/10.1145/3205455.3205539>
   4. Urbanowicz, R. J., Meeker, M., **LaCava, W.**, Olson, R. S., & Moore, J. H. (2018). Relief-based feature selection: introduction and review. *Journal of Biomedical Informatics.* In Press. arXiv preprint arXiv:1711.08421

**Total citations: 994; h-index: 15; i10-index: 27 (source: Google Scholar)**

**NCBI biobliography:** [**https://www.ncbi.nlm.nih.gov/myncbi/collections/bibliography/53476482/**](https://www.ncbi.nlm.nih.gov/myncbi/collections/bibliography/53476482/)

**D. Research Support and/or Scholastic Performance**

**Active**

**K99/R00**  LM012926-01A1 La Cava (PI) 06/01/19 -

NIH/NLM

Multi-objective representation learning methods for interpretable predictions of patient outcomes using electronic health records

Role: PI

Completed Research Support

Postdoctoral Fellowship Warren Center for Network and Data Sciences 08/08/16 – 08/08/17

School of Engineering and Applied Science, University of Pennsylvania

Feature learning for biomedical data

This fellowship supports interdisciplinary collaboration between mentors and fellows to support cutting edge data science and network science.

Role: Fellow

NSF IGERT Fellowship University of Massachusetts Amherst 09/01/2012 - 05/31/2016

UMass IGERT Offshore Wind Energy Program

The goal of this Integrative Graduate Education and Research Traineeship program was to create a community of researchers who understand the technological challenges, environmental implications, and socioeconomic and regulatory hurdles of offshore wind farms.

Role: Fellow

TG-ASC140043 La Cava (PI) 2014-2015

Extreme Science and Engienering Discovery Environment (XSEDE)

Automatic Identification of Dynamic Models for Complex Systems

This award provided high performance computing support for large-scale machine learning.

Role: PI