Citation:
T. Adali, R. C. Guido, T. K. Ho, K. -R. Müller and S. Strother, "Interpretability, Reproducibility, and Replicability [From the Guest Editors]," in IEEE Signal Processing Magazine, vol. 39, no. 4, pp. 5-7, July 2022, doi: 10.1109/MSP.2022.3170665.

DOI:
https://doi.org/10.1109/MSP.2022.3170665

# Introduction for the Special Issue on Explainability in Data Science: Interpretability, Reproducibility, and Replicability

Tülay Adali, Rodrigo Capobianco Guido, Tin Kam Ho, Klaus-Robert Müller, and Stephen Strother

Most of the work that we do in signal processing these days is data driven. The shift from the more traditional and model-driven approaches to those that are data driven also underlined the importance of explainability of our solutions. Because most traditional signal processing approaches start with a number of modeling assumptions, they are comprehensible by the very nature of their construction. However, this is not necessarily the case when we choose to rely more heavily on the data and minimize modeling assumptions.

Explainability is critical not only for the simple reason that one would like to have confidence over the solutions, but also because one would like to obtain further insights about the problem from the learned models. This includes interpretability and completeness so that one can not only “audit” them, but also ask appropriate questions to probe for insights beyond the initial solution, and address additional concerns such as safety, fairness, and reliability. Interpretability, i.e., ability to attach a physical meaning to the solution, along with reproducibility, and replicability are three key aspects of explainability. Following the definitions by the National Academies of Sciences, Engineering, and Medicine, reproducibility refers to obtaining consistent results using the same data and code—i.e., method,—as the original study, and replicability is obtaining consistent results across studies aimed at answering the same scientific question using new data or other computational methods.

In this special issue we have nine articles that demonstrate the multi-faceted nature of explainability, and span the related concepts of interpretability, reproducibility, and replicability. They successfully demonstrate the rich nature of these concepts, while also highlighting the fact that they take on slightly different meanings in different contexts, and the considerations might be slightly different as well. These papers also underline the fact that explainability is a key theme that requires attention across different application domains and types of solutions, i.e., well beyond neural networks where they have been mostly emphasized to date.

The first two papers of the special issue study the questions of reproducibility, replicability and interpretability for two important classes of machine learning solutions, matrix and tensor decompositions (MTDs) and graph data science. The first article "Reproducibility in Matrix and Tensor Decompositions” by Adali et al. addresses reproducibility in MTD solutions that have been growing in importance, where in addition to the discovery of structure in the data, the resulting decomposition is also directly interpretable. With an applied focus where there is no ground truth, authors study the intricate relationships of interpretability, model match, and uniqueness. They make use of two widely used methods with relaxed uniqueness guarantees, independent component analysis and the canonical-polyadic decomposition, and provide examples to solidify these concepts and demonstrate the tradeoffs. Finally, a reproducibility checklist for MTDs is provided similar to those developed for supervised learning. In “Explainability in Graph Data Science”, Aviyente and Karaaslanli explain methods and metrics from network science to quantify three different aspects of explainability, i.e., interpretability, replicability and reproducibility, in the context of community detection. Specifically, the strategies described by the authors can be used to address some common issues and provide guidelines to reduce the opacity of community detection algorithms and their outputs. In addition, they can be extended to other community detection and data clustering algorithms as well as different learning tasks on graphs.

The second group of papers take a broader view of explainability with a focus on neural networks. Letzgus et al. discuss an area of explainable artificial intelligence (XAI) that has so far received comparatively little attention, namely XAI for regression models. Their review, "Toward Explainable Artificial Intelligence for Regression Models," provides novel theory showing that there are important conceptual differences between XAI for regression and classification, not only algorithmically but also with respect to the choice of reference for the explanation. The paper "Explanatory Paradigms in Neural Networks", by AlRegib and Prabhushankar characterizes a complete explanation as an additive combination of observed correlation, counterfactual and contrastive explanations. It then discusses how existing explanation methods can be analyzed within this framework, and how well they are suited to different evaluation strategies under a proposed taxonomy.

The next two papers consider generative adversarial networks, which have been growing in importance. In “Robust Explainability" Nielsen et al. present a timely and comprehensive tutorial on gradient-based attribution/salience methods, their relationship to adversarial robustness and the practical importance of robust explainability of computer vision classification models based on these techniques, together with many of the associated terms that appear in the explainability literature. They provide a useful list of best practices to consider when choosing an explainability method and conclude with future directions of research in the area of robust explainability. They augment their paper with a website with links to all the explainability methods discussed, the paper's figures, and the code for generating the figures. In the second paper of this group “Explaining Artificial Intelligence Generation and Creativity,” Das and Varshney review different motivations, algorithms, and methods intended to explain the principles of AI algorithms or the possible artifacts they produce, by using a generative point-of-view with creativity as the focus. Particularly, they observe that discussions of interpretable AI, especially in settings of decisions and predictions, frequently start with the

misconception that there is a fundamental trade-off between interpretability and accuracy, however, as reviewed, numerous examples show the opposite.

The last group of papers of the special issue consider explainability with an application focus, and across a wide array of data-driven solutions. The first two papers consider applications in healthcare and the last one on time series classification. The first paper of this group, “Explainability of Methods for Critical Information Extraction From Clinical Documents”, by Ho, Luo, and Guido reviews a collection of representative works that address several natural language understanding tasks in healthcare, and discusses their explainability. It showcases the complex dimensions of considerations in providing explainable methods for an essential application of artificial intelligence. In “Interpreting Brain Markers” Jiang et al. review predictive methods and their applications for interpreting brain signatures in neuroimaging based on a survey of more than 300 articles. This allows to better validate and assess reliability and interpretability of biomarkers across multiple datasets and contexts. Finally, in "Post Hoc Explainability for Time Series Classification," Mochaourab et al. discuss the explainability advantages of methods for time series classification that are based on representations well established in signal processing. The paper highlights the relevance of such conventional transformations for the important concerns of understanding feature importance and providing counterfactual explanations.

We thank our contributors for their comprehensive and interesting articles, Robert Heath for his support for our proposal, and Christian Jutten for providing valuable guidance and support at every step of the process. We would like to also extend our thanks to our reviewers for their detailed and insightful comments, to Rebecca Wollman for the guidance and support along the way, and to Sharon Turk for the special care in putting together our special issue.

Data-driven solutions are becoming the dominant approach in practical problems across many domains in the sciences and technology, and explainability is a key aspect that will further enhance their utility. Signal processing is at the heart of data science, and is where the connection with applications is natural. Hence, we are hoping the insights as well as the critical perspectives provided by the contributions in our special issue will prove to be a useful reference and will help identify some of the new and emerging directions in the area.

## Guest Editors

***Tülay Adali*** (adali@umbc.edu) received her Ph.D. degree in electrical engineering from North Carolina State University. She is a distinguished university professor in the Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County, Baltimore, Maryland, 21250, USA. She is a past vice president for technical directions for the IEEE Signal Processing Society (SPS) and is currently the chair-elect of IEEE Brain. She is a Fellow of IEEE and the American Institute for Medical and Biological Engineering, a Fulbright Scholar, and an SPS Distinguished Lecturer. She is the recipient of a Humboldt Research Award, an IEEE SPS Best Paper Award, the University System of Maryland Regents’ Award for Research, and a National Science Foundation CAREER Award. Her research interests include statistical signal processing and machine learning and their applications, with emphasis on applications in medical image analysis and fusion.

**Rodrigo Capobianco Guido** (guido@ieee.org) received his Ph.D. degree in computational applied physics from the University of São Paulo (USP), Brazil, in 2003. After two postdoctoral programs in signal processing at USP, he obtained the title of associate professor (livre-docência) in signal processing, also from USP, in 2008. Currently, he is an associate professor at São Paulo State University (UNESP) at São José do Rio Preto, São Paulo, 15054-000, Brazil. Guido has been an area editor for IEEE Signal Processing Magazine and, recently, has been included in Stanford University ranking of the world’s top 2% scientists. He is a Senior Member of IEEE.

***Tin Kam Ho*** (tkh@ieee.org) received her Ph.D. degree in computer science from the State University of New York at Buffalo in 1992. She is a senior artificial intelligence scientist at IBM Watson Health, Yorktown Heights, New York, 10598-0218, USA, where she leads projects in semantic modeling of natural languages in clinical applications. Prior to 2014, she was with Bell Labs in Murray Hill as the head of the Statistics and Learning Research Department. She pioneered research in multiple classifier systems and ensemble learning, random decision forests, and data complexity analysis and also contributed to many applications of pattern recognition and computational modeling. Ho is a Fellow of IEEE and the International Association for Pattern Recognition.

***Klaus-Robert Müller*** studied physics at the Technische Universität Karlsruhe, Karlsruhe, Germany and received the Ph.D. degree in computer science from there in 1992. He has been a Professor of computer science with TU Berlin, since 2006. From 2012 he has at the same time been a distinguished Professor at Korea University. In 2020 and 2021, he was on a sabbatical leave from TU Berlin and with the Brain Team, Google Research, Berlin, Germany as a Principal Researcher. He is directing the Berlin Institute for the Foundations of Learning and Data (BIFOLD). Dr. Müller was elected member of the German National Academy of Sciences, Leopoldina, in 2012 and the Berlin Brandenburg Academy of Sciences in 2017 and the National Academy of Science and Engineering in 2021, and an External Scientific Member of the Max Planck Society in 2017. He is the lead of the ELLIS unit Berlin. From 2019, he became an ISI Highly Cited Researcher in the cross-disciplinary area. Among others, he was awarded the Olympus Prize for Pattern Recognition in 1999, the SEL Alcatel Communication Award in 2006, the Science Prize of Berlin by the Governing Mayor of Berlin in 2014, the Vodafone Innovations Award in 2017, and the 2020 Best Paper Award in the journal Pattern Recognition. His research interests are intelligent data analysis and Machine Learning in the sciences (Neuroscience (specifically Brain-Computer Interfaces, Physics, Chemistry) and in industry.

***Stephen Strother*** (sstrother@research.baycrest.org) received his Ph.D. degree in electrical engineering in 1986 from McGill University, Montreal, Canada. He is a member of the Rotman Research Institute, Toronto, M6A 2E1, Canada, and a professor of medical biophysics at the University of Toronto. His research interests include neuroinformatics and data science for neuroimaging and big clinical data sets through statistical and machine learning techniques, applying these techniques in cognitive neuroscience and brain disease and translating this work to nonacademic settings.