The Journal of Open Source Education

# What and How of Machine Learning Transparency: Building Bespoke Explainability Tools with Interoperable Algorithmic Components

**Kacper Sokol**[1, 2], **Alexander Hepburn**[1], **Raul Santos-Rodriguez**[1], **and Peter Flach**[1]

**1** Intelligent Systems Laboratory, University of Bristol, United Kingdom **2** ARC Centre of Excellence for Automated Decision-Making and Society, RMIT University, Australia

**DOI:** 10.21105/jose.00175

**Software**

- Review
- Repository
- Archive

**Submitted:** 09 April 2022
**Published:** 21 October 2022



## Summary

Explainability techniques for data-driven predictive models based on artificial intelligence and machine learning algorithms allow us to better understand the operation of such systems and help to hold them accountable (Sokol & Flach, 2021a). New transparency approaches are developed at breakneck speed, enabling us to peek inside these black boxes and interpret their decisions. Many of these techniques are introduced as monolithic tools, giving the impression of one-size-fits-all and end-to-end algorithms with limited customisability. Nevertheless, such approaches are often composed of multiple interchangeable modules that need to be tuned to the problem at hand to produce meaningful explanations (Sokol et al., 2019). This paper introduces a collection of hands-on training materials – slides, video recordings and Jupyter Notebooks – that provide guidance through the process of building and evaluating bespoke modular surrogate explainers for tabular data. These resources cover the three core building blocks of this technique: interpretable representation composition, data sampling and explanation generation (Sokol et al., 2019).

## Modular Surrogate Explainers

The training materials introduce the concept of *modular* explainability algorithms using the example of surrogate explainers for tabular data. This separation of functionally independent building blocks allows us to consider the influence of each component, and their interdependence, on the robustness and faithfulness of the final explainer. To this end, we review a collection of techniques to evaluate the quality of the modules and their overall effectiveness. These metrics can guide the parameterisation of the entire explainability algorithm, providing an opportunity to tune it to the problem at hand. All of these insights demonstrate that while surrogate explainers are model-agnostic and post-hoc – i.e., they work with any black box and can be retrofitted into pre-existing predictive models, thus making them a popular choice for explaining black-box predictions (Ribeiro et al., 2016) – using off-the-shelf explainability approaches may result in subpar performance for individual use cases (Rudin, 2019). Therefore, understanding how to build a bespoke surrogate explainer that is suitable for a particular situation is a prerequisite for trustworthy and meaningful explainability of data-driven systems and their decisions.

Prior to diving into the practicalities of composing surrogate explainers, the training materials introduce the concept of algorithmic explainability of predictive models and

The Journal of Open Source Education

discuss the fundamental ideas behind surrogates for text, image and tabular data. This theoretical overview is followed by a brief presentation of the software used for the hands-on modules; `FAT Forensics`[1] is an open source Python package designed for inspecting selected fairness, accountability and *transparency* aspects of data (and their features), *models* and *predictions* (Sokol et al., 2020, 2022). Having covered the basics, the practical coding resources focus on the three building blocks of surrogate explainers for tabular data identified by the bLIMEy – build LIME yourself (Sokol et al., 2019) – meta-algorithm:

- interpretable (data) representation composition;
- data sampling; and
- explanation generation (interpretable feature selection, data sample weighting, surrogate model training and explanation extraction).

These learning modules review some of the interoperable algorithmic components available at each step, discuss their pros and cons for a range of applications, guide through their optimal selection strategies and propose suitable evaluation criteria – see Figure 1. In particular, interpretable representations are built with quartile discretisation and decision trees (Sokol & Flach, 2020b); data are generated with Gaussian and mixup sampling (Sokol et al., 2019; Zhang et al., 2018); and explanations are extracted from linear and tree-based surrogate models (Sokol & Flach, 2020a). Notably, these choices determine the type, role and quality of the resulting explanations composed for black-box predictions. Therefore, these hands-on materials illustrate how such interoperable algorithmic building blocks behave in various scenarios and demonstrate how to use these components to configure robust explainers with well-known properties and failure modes based on first-hand observations and a collection of quantitative evaluation metrics and validation techniques.

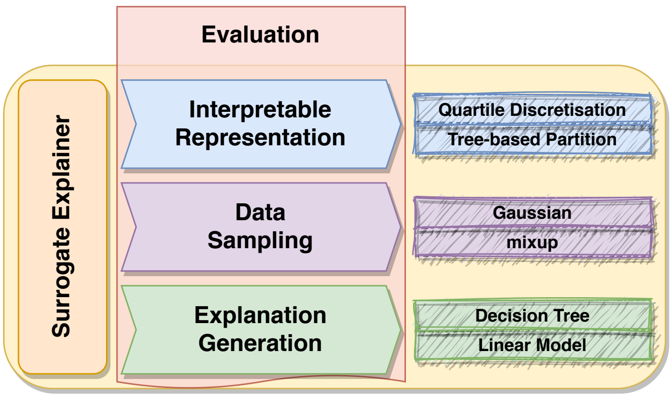


**Figure 1:** Overview of surrogate explainers modularity listing components specific to tabular data.

The introduction to algorithmic explainability; the theoretical overview of surrogate explainers for text, image and tabular data; and the outline of `FAT Forensics` are presented in a collection of *slides* and *instructional video recordings*. The hands-on materials are delivered with *Jupyter Notebooks* that interweave textual guidance, code examples and informative plots. All the insights learnt throughout the practical exercises enable the tutees to create robust surrogate explainers for an arbitrary black-box predictive model built for their own tabular data set. The training resources are designed to appeal and be accessible to an audience with a wide range of backgrounds and experiences. Active participation in the practical part requires basic familiarity with Python programming and access to a computer connected to the Internet, which enables execution of the Jupyter Notebooks online without installing any software on a personal machine.

[1] https://fat-forensics.org/

The Journal of Open Source Education

These training materials were used to deliver a hands-on tutorial – of the same title – at the 2020 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)[2], the recordings of which are available on YouTube[3]. Moreover, they inspired a number of interactive sessions at various summer schools aimed at doctoral students in artificial intelligence and machine learning, as well as undergraduate lectures, academic presentations and invited talks. The slides, extra hands-on resources and video recordings of some of these events are available on the FAT Forensics Events website[4]. The new teaching materials[5] additionally cover surrogates for image data – focusing on the influence of segmentation granularity and occlusion colour on the trustworthiness of the resulting explanations (Sokol & Flach, 2020b) – and touch upon other explainers such as *permutation importance* (Breiman, 2001), *individual conditional expectation* (Goldstein et al., 2015) and *partial dependence* (Friedman, 2001).

Notably, these sessions propelled the improvement, evolution and expansion of the training resources. In particular, the programming-focused exercises have been complemented by *no-code* Jupyter Notebooks that enable interactive experimentation with the explainers through intuitive *Jupyter Widgets*, thus allowing to better engage with the audience in a limited time. The same strategy has been employed for the slides – by embedding interactive examples based on widgets – to which end they have been built with RISE[6]. From our experience, the teaching became much more effective when the ubiquitous PDF slides and Jupyter Notebook programming exercises were replaced with and/or enriched by formats supporting seamless interaction with the taught material (in our case achieved through widgets). This exploration of alternative technologies for building training resources has also inspired a prototype of a new publishing workflow, where multiple artefacts such as online documents, slides and computational notebooks can be composed from a unified collection of source materials (Sokol & Flach, 2021b).

# Statement of Need

The training resources described by this paper introduce a novel learning paradigm for algorithmic explainability of data-driven predictive systems based on artificial intelligence and machine learning techniques. Instead of treating these tools as end-to-end, monolithic entities whose configuration is only facilitated through the parameters exposed by their developers, these educational materials look into their modularity to identify atomic and functionally interoperable building blocks. By decomposing explainers into their core elements we can better understand their role and configure them for the application at hand. Within this purview, such techniques are diagnostic tools that *only* become explainers when their properties and interpretation of their outputs are well understood and designed accordingly. Therefore, to engender trust in data-driven predictive systems, the employed explainability approaches must be trustworthy themselves in the first place – the learning objective underlying the interactive coding exercises. The training materials achieve these goals by supporting the following learning outcomes specifically for surrogate explainers (which were chosen because of their flexibility and popularity):

- identify self-contained algorithmic components of the explainer and understand their functions;
- connect these building blocks to the explainability requirements unique to the investigated predictive system;
- select appropriate algorithmic components and tune them to the problem at hand;

[2] https://events.fat-forensics.org/2020_ecml-pkdd/
[3] https://www.youtube.com/playlist?list=PLgdhPOmeUNm0H2XTQECK3wabnDohZURLK
[4] https://events.fat-forensics.org/
[5] https://github.com/fat-forensics/resources/
[6] https://rise.readthedocs.io/

- evaluate these building blocks (in this specific context) independently and when joined together to form the final explainer; and
- interpret the resulting explanations in view of the uncovered properties and limitations of the bespoke explainability algorithm.

The modularity and diversity of these training materials – slides, video recordings and Jupyter Notebooks – allow them to be adapted or directly incorporated into a course on explainable artificial intelligence and interpretable machine learning, or form the basis of a range of educational resources such as practical training sessions and conference tutorials. The module can be taught as is – reusing the slides and computational notebooks – with either bespoke tuition or by following the prerecorded videos. Alternatively, the narration, figures and results presented within the notebooks may be shaped into tailor-made teaching materials. The hands-on resources can also become a standalone case study supplementing relevant explainability and interpretability courses. The comprehensive and in-detail presentation of the topic, covering both the underlying theory and practical aspects, is suitable for and accessible to undergraduate and postgraduate students, researchers as well as engineers and data scientists interested in the subject. This module fills a gap in educational materials dealing with artificial intelligence and machine learning transparency by focusing on understanding of the inner workings of these techniques and the influence of their building blocks on the robustness, veracity and comprehensibility of explanatory insights into black-box predictive models.

## Acknowledgements

This work was partially supported by TAILOR (Trustworthy AI through Integrating Learning, Optimisation and Reasoning), a project funded by EU Horizon 2020 research and innovation programme under GA No 952215; the UKRI Turing AI Fellowship EP/V024817/1; and the ARC Centre of Excellence for Automated Decision-Making and Society, funded by the Australian Government through the Australian Research Council (project number CE200100005).

## References

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232. https://doi.org/10.1214/aos/1013203451

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, *24*(1), 44–65. https://doi.org/10.1080/10618600.2014.907095

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Sokol, K., & Flach, P. (2020a). LIMEtree: Interactively customisable explanations based on local surrogate multi-output regression trees. *arXiv Preprint arXiv:2005.01427*. https://doi.org/10.48550/arXiv.2005.01427

Sokol, K., & Flach, P. (2020b). Towards faithful and meaningful interpretable representations. *arXiv Preprint arXiv:2008.07007*. https://doi.org/10.48550/arXiv.2008.07007

Sokol, K., & Flach, P. (2021a). Explainability is in the mind of the beholder: Establishing the foundations of explainable artificial intelligence. *arXiv Preprint arXiv:2112.14466*. https://doi.org/10.48550/arXiv.2112.14466

Sokol, K., & Flach, P. (2021b). You only write thrice: Creating documents, computational notebooks and presentations from a single source. *Beyond Static Papers: Rethinking How We Share Scientific Understanding in Machine Learning – ICLR 2021 Workshop*. https://doi.org/10.48550/arXiv.2107.06639

Sokol, K., Hepburn, A., Poyiadzi, R., Clifford, M., Santos-Rodriguez, R., & Flach, P. (2020). FAT Forensics: A Python toolbox for implementing and deploying fairness, accountability and transparency algorithms in predictive systems. *Journal of Open Source Software*, *5*(49), 1904. https://doi.org/10.21105/joss.01904

Sokol, K., Hepburn, A., Santos-Rodriguez, R., & Flach, P. (2019). bLIMEy: Surrogate prediction explanations beyond LIME. *Workshop on Human-Centric Machine Learning (HCML 2019) at the 33rd Conference on Neural Information Processing Systems (NeurIPS)*. https://doi.org/10.48550/arXiv.1910.13016

Sokol, K., Santos-Rodriguez, R., & Flach, P. (2022). FAT Forensics: A Python toolbox for algorithmic fairness, accountability and transparency. *Software Impacts*, *14*, 100406. https://doi.org/10.1016/j.simpa.2022.100406

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). mixup: Beyond empirical risk minimization. *Proceedings of the 6th International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.1710.09412