Novel Topological Shapes of Model Interpretability

Hendrik Jacob van Veen
MLWave
info@mlwave.com

Abstract

The most accurate models can be the most challenging to interpret. This paper advances interpretability analysis by combining insights from `Mapper` with recent interpretable machine-learning research. Enforcing new visualization constraints on `Mapper`, we produce a globally - to locally interpretable visualization of the Explainable Boosting Machine. We demonstrate the usefulness of our approach to three data sets: cervical cancer risk, propaganda Tweets, and a loan default data set that was artificially hardened with severe concept drift.

1 Introduction

The question of how to optimally generalize from train data to test data is quintessential to research in machine learning (ML) [48]. In practice, accuracy is hardly a sole optimality concern. There are other desiderata, such as simplicity, fairness, and robustness to concept drift. Even with the focus on accuracy, complex modern ML algorithms will misuse artifacts in the data to improve the accuracy on the train set, thereby falsely suggesting better performance than should be expected in reality. For instance, Caruana et al. [8] demonstrated that a hospital-triage model learned to attribute *low*-risk to asthma patients, due to the aggressive – and effective – care such *high*-risk asthma patients had received in the train data. Financial risk models may learn from – and become entwined with – business artifacts [4], or overfit to seasonality, causing financial ruin after a crash [12]. Protected features are automatically inferred from other features [35]. Even benchmark-winning computer-vision models display unpredictable failure modes, reducing trust in high-stakes deployments [17].

Interpretability is therefore important. In some cases it is even required by law [41]. Trust in models increases with interpretability analysis and input from domain experts, because this can correct mistakes. Ideally, interpretability not only helps with finding bias and debugging, but can also be used to inform policy and research [15]. Unfortunately, a trade-off between accuracy and other desiderata has been observed. Black-box algorithms, such as deep neural networks and gradient boosted decision trees [9], produce highly accurate predictions and thus see a wide adoption in industry and ML competitions, yet such models are not very interpretable. Conversely, white-box algorithms, such as logistic regression and decision rules on handcrafted features, are more interpretable, but can fall behind in accuracy on the problems that modern-day ML practitioners face.

The ML community has addressed accuracy trade-offs in a number of ways. Some relieve tension by separating accuracy from other concerns, such as interpretability or fairness. For instance, LIME and SHAP use white-box models and game theory to explain black-box predictions [30], while it is also possible to threshold the predictions of a black box model [19] to introduce fairness constraints [2]. Turner [45] weakens interpretability requirements and Rolf et al. [37] jointly optimizes for performance (profit) and society-aware cost functions. One may even question the existence of trade-offs [38]. Finally, a promising approach is to create inherently high-accuracy highly-interpretable models, as in the case of the Explainable Boosting Machine (EBM) from InterpretML [33]. Here, we examine the EBM in combination with the `Mapper` [43] from Topological Data Analysis (TDA) [6].

the Topological Data Analysis and Beyond Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Contributions. In this paper, we leverage the strengths of `Mapper` and interpretable ML to construct a new method of interest to practitioners and researchers in ML and TDA. For interpretable ML we (1) offer a more holistic data analysis, allowing analysts to go from a compressed chart to a single sample, (2) create cluster explanations by aggregating local explanations, and (3) show interpretable clusters with persistence through time. For `Mapper`, we (4) place `Mapper` in the context of traditional analysis, potentially easing adoption, and (5) make `Mapper`'s output carry more useful meaning than it currently has. We (6) show the utility and scalability of this approach on three challenging data sets.

2 Methods and Experiments

Hypothesis. The EBM rivals state-of-the-art models and is very interpretable. The `Mapper` from TDA allows for unsupervised data exploration. Thus, we hypothesize that extending the EBM with `Mapper` techniques – combining global and local explanations in a single interface – allows for more informative visualizations than charts alone. In the next paragraphs we briefly describe the algorithmic details of the EBM and `Mapper` as these relate to our methods.

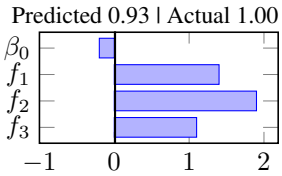


Figure 1: An EBM scores a local sample explanation.

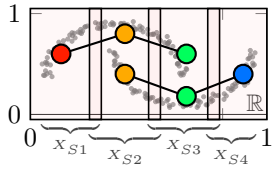


Figure 2: `Mapper` covers data with overlapping bins. Then it clusters inside these bins and connects all the clusters that have points in common.

The **EBM** is based on Generalized Additive Models (GAM) [20] and is of the form $g(E[y]) = \beta_0 + \sum f_j(x_j)$ with g a link function to switch between regression and classification. The EBM improves on the GAM with feature interactions, bagging, and boosting [26]. A prediction is obtained by simply summing the scores of features. So, it is natural to create local sample explanations and global feature explanations and to subsequently chart such explanations for interpretability analysis. However, analysts are not currently able to quickly switch between global - and local explanations, and hence the contextual value of feature explanations may suffer. Figure 1 shows a local explanation.

`Mapper` is an algorithm to visualize the data created in TDA with network graphs. It draws inspiration from the Reeb graph [1]. Intuitively, `Mapper` works by a local clustering stratified by a function, such as `HEIGHT(`X`)`. More precisely, start with a filter function f (also called a "lens") and project X down to one dimension $\mathbb{R}$; $f : X \to \mathbb{R}$. Construct an overlapping covering of $\mathbb{R}$. From each i intervals on $\mathbb{R}$, collect i subsets from X; $X_{S1}, X_{S2}, \ldots, X_{Si}$. Create clusters inside each subset X_S with clusterer $\mathcal{C}$ to form vertices V. Finally, edges E are created by drawing an edge E_{ij} between vertices V_i and V_j when these share data points $V_i \cap V_j \neq \emptyset$. The output is a graph $\mathcal{G}(V, E)$ which can be displayed on screen for analysis. See Figure 2 for an example. `Mapper` is a versatile approach to capturing shapes of data, in part due to the flexibility of the parameters. Hence, it sees applications to problems in a range of academic fields, such as healthcare [39], security [10], and neuroscience [16]. Moreover, `Mapper` offers unique insights to analysts, and provides hypothesis that scale with large amounts of complex data [27]. Even so, one needs basic knowledge of algebraic topology to effectively use TDA, reducing utility for non-experts. Furthermore, since the position of a node on the screen has no contextual meaning, analysts need extra steps to place areas of interest into context, for instance by looking at the descriptive statistics for a node. Figure 3 illustrates how we solve this.

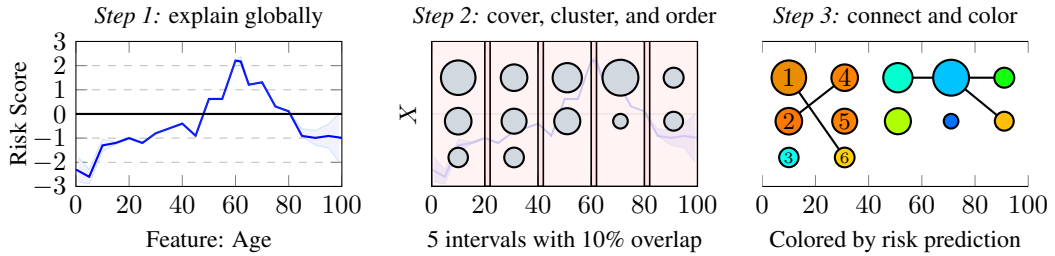


Figure 3: **Illustration of part of our approach.** In *Step 1* a chart with a global feature explanation is created. In *Step 2* we map this chart on the x-axis and order clusters by membership size. In *Step 3*, the clusters are connected and colored by a function of interest. Note how, unlike traditional `Mapper`, while Clusters 1 and 2 are not connected, these are still close-by on the screen. Note also how the edge between Clusters 1 and 6 shows a persistent – but downward – trend. Finally, note how the color of Cluster 3 surfaces a small *high*-risk group in a *low*-risk feature-value interval (Age < 23).

Setup. Following Algorithm 1 (in Appendix A), we apply `Mapper` to traditional charts. In our case, we use a lens for each axis of a traditional chart to output two topological graphs. The horizontal axis is a continuous variable, such as a time tick, or an age column. The vertical axis is the output from a model, either an interpretation score, or an opaque dynamical system, such as the stock market. Where relevant, we limit the number of intervals per axis in an effort to reduce cognitive load [29]. Unlike traditional `Mapper` methods, the nodes are confined to the intervals in which these were formed and the clusters inside an interval are ordered by set size. We show that these simple modifications creates more insightful data analysis, without removing any of `Mapper`'s benefits or theoretical grounding. We perform clustering with DBSCAN [13], which is density-based clustering, motivated by the speed and not being required to set the number of clusters, and agglomerative clustering [47], motivated by its strong theoretical guarantees and past usage in literature [32]. By coloring nodes with the average predictions in a cluster, we automatically surface low-score clusters inside high-score intervals. For the local sample explanations, we order the features by impact (due to space limitations we only show the top 4). Since the standard EBM implementation provides a Scikit-Learn API [5] we look at `Mapper` implementations with a similar API, such as giotto-tda [44] and KeplerMapper [46], and we opt for the latter. Appendix A has more details on the experiments.

Data. To demonstrate potential, we use a healthcare data set [14], motivated by the importance of interpretability in the field of healthcare, and by the prospect of informing disease treatment and risk factors [31]. After pre-processing, the data have 849 rows and 32 features. We do a stratified split of 20%. We use a social media propaganda data set [36] to place our algorithm in the context of a timely problem of adverse shaping of political discourse and to demonstrate that our approach also works on unstructured text data, and without the EBM. The data set has 120.058 rows and 3.000 features. Our final choice of data set is motivated by the tricky real-world issues implementing credit-risk models. Since such data is generally not available, we altered a public loan data set [49] to display very problematic concept drift (different feature distribution in train and test set). We use 25.000 samples for training and 5.000 samples for testing. The first 2.500 test samples are regular samples, the second 2.500 test samples are corrupted by randomly shuffling 50% of the features in each column. Appendix B has reproducibility checklists for all data.

Models and parameter settings. We use the EBM from InterpretML. For benchmarks in the cervical cancer data set and the loan data set, we compare with a non-tuned XGBoost [9], for that is a known accurate model. Since a thorough performance comparison is not the point of this paper, we do not hypertune, but manually change parameters to make the EBM approximately equal XGBoost. The propaganda tweets data set is created by taking char-grams of length four to six and calculating TF-IDF for the top 3000 features. For comparisons see Appendix C and for full EBM vs. XGBoost benchmark details see the InterpretML repository [28].

Discussion of results. Figure 4 demonstrates how the current state-of-the-art in model interpretability can be improved with our `Mapper` approach. Notice how analysts are able to go from a global feature explanation, to a cluster explanation, to a local sample explanation, all from within a single interface. Figure 5 shows our approach applies to the challenging real-life problem of concept drift and demonstrates its utility for debugging a deployed ML model. The color and connectivity of the `Mapper` graph hint at a distribution change, thereby pinpointing interesting areas for further analysis.

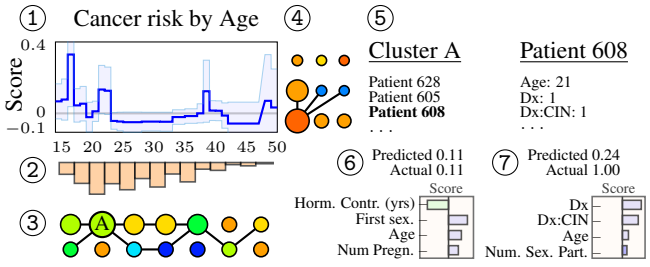


Figure 4: Resulting **interface for a cervical cancer data set.** ① and ② are created by the EBM. ③ and ④ are the `Mapper` graphs, positioned through confining and ordering. ⑤ shows the members in the selected cluster. ⑥ is the averaged cluster explanation and ⑦ the local explanation of a selected member.

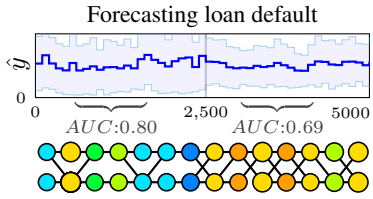


Figure 5: Result of **concept drift experiment.** Clusters with concept drift (in the right region) are more connected. The colors in both regions are different. X-Axis are row # and Y-axis is mean(prediction $\hat{y}$ per 100 rows). $\mathcal{C}$ is agglomerative clustering. Color is stdev($\hat{y}_{100}$).

Figure 6 indicates that our approach is fitted to unstructured data, such as text. Our method generates interesting questions, which our visualization can instantly answer, such as: *"What was the longest-lasting topic?"*, *"Which topic is most atypical?"*, and *"Which topics correlate with increased activity?"*. Remarkably, the number of clusters seems to encode for density of topics in an interval.

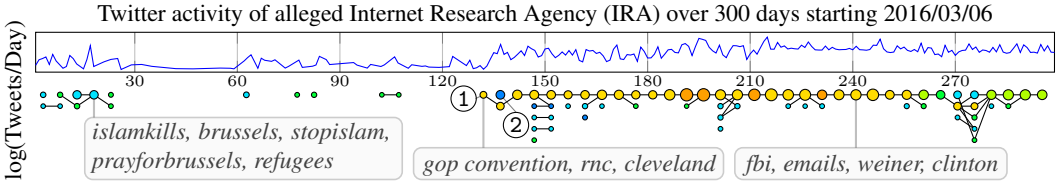


Figure 6: The result of `Mapper` on **alleged IRA tweets.** We apply clustering with DBSCAN on X, which is a TF-IDF representation [21] with 3.000 columns (formed from the top char-grams of length 4 to 6) and 120.058 rows. Colored by Cosine distance to mean(X). Cluster ① signals start of longest-lasting topic chain. Cluster ② is most distant to mean.

3 Conclusions

We showed how extending the interpretability research of ML with the `Mapper` from TDA improves data analysis and model interpretability. We demonstrated that our approach – laying out network graphs alongside a traditional chart – is able to capture novel shapes of model interpretability which persist (or not) over time. We showed that the local clustering of `Mapper` finds clusters of interest, that traditional interpretability methods cannot, for instance, low-score clusters inside high-score regions, and thus is a natural extension to the process of making models less biased and unfair. By averaging the local explanations inside of a cluster we obtained a new type of aggregate explanation useful for data - and cluster analysis. We favorably examined usefulness on three data sets.

Relation to other work. Cluster analysis is not new [18]. Two decades ago, Nauck et al. advocated interpretable models over black-box models in high-stakes settings [31]. Olah et al. explored the power of combining different interpretability techniques [34]. Following these trends, we also find that combining local sample explanations with global feature explanations provides a more powerful interface. More recently, Carlsson et al. showed that using the output of a ML model as a filter function yields valuable insights into failure modes [7]. Saul et al. applied `Mapper` to the predictions of a black-box model to create global and local explanations [40]. In this regard, the closest related work we found in functionality in KeplerMapper (aggregate cluster statistics) and an article by Harlan Sexton [42] which describes averaging the images within a cluster. Our approach builds on `Mapper` [43]. `Mapper` on original features is also not uncommon. However, unlike pure `Mapper` output, the position of a node on the screen now has a useful meaning. Since we also order the clusters inside an interval by membership size, the angle of the edges now carries meaning too, with horizontal edges connecting clusters of similar size. The "branching out" of text topics offers a venue for future research, representing complex objects, such as literature – or indeed an entire era [3] – with graphs.

Impacts and limitations. Since accurate *and* interpretable modeling removes an economic incentive to favor profit over interpretation, we expect our solution to contribute to a more insightful and more responsible application of ML systems in society. The authors do appreciate that interpretable ML is a dual-use technology (*"How do we get at-risk life-insurance cohorts to cancel?"*) and that fair models may be deployed in unfair settings. For completeness, we highlight three possible negative impacts and shortcomings of our work. First off, our approach currently has no open-source implementation and we did not provide a unified framework for parameter settings and lens choice. Therefore, replicability is hampered, and an objective explanation of a model is impossible. Second, interpretability and fairness is not a solved problem in theory nor practice. Though using a surrogate model to explain a closed-box model may improve interpretability, there is no guarantee that these models reasoned exactly the same [30]. Additionally, fairness and interpretability is ill-defined [25], may ignore essential elements of causality [11], and simultaneously satisfying all fairness constraints is provably impossible [23]. Thus, strongly implying the creation of a fair or debiased model using our approaches may be deceptive and create a false sense of fairness. Lastly, creating fair and safe models requires the indispensable input of domain experts. Topological data analysis may not be the best medium for non-experts in TDA and hence reduce the quality of their feedback [24]. This third impact inspired the combination of traditional charts and topological graphs, instead of outputting

the topological graphs alone, but work remains to be done. Though outside the scope of this paper, solving for issues with low-error implementation, accessibility, and user misalignment [22] is a must.

Acknowledgements

The authors would like to thank Nicholas Domene for helpful discussions to improve clarity and Nathaniel Saul for help with structuring and mock reviewing. Furthermore, we are thankful to the anonymous reviewers for their thorough and helpful suggestions and to the organizers of the Topological Data Analysis and Beyond Workshop.

References

[1] Henry Adams. The mapper algorithm and reeb graphs. *Applied Algebraic Topology Research Network*, 2020. https://youtu.be/_tiv0qYcM3U.

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. volume 80 of *Proceedings of Machine Learning Research*, pages 60–69, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[3] Alexander TJ Barron, Jenny Huang, Rebecca L Spang, and Simon DeDeo. Individuals, institutions, and innovation in the debates of the french revolution. *Proceedings of the National Academy of Sciences*, 115(18):4607–4612, 2018.

[4] Leon Bottou. Two big challenges in machine learning. In *Proceedings from 32nd International Conference on Machine Learning*, 2015.

[5] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, et al. Api design for machine learning software: experiences from the scikit-learn project. 2013.

[6] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[7] Leo S Carlsson, Mikael Vejdemo-Johansson, Gunnar Carlsson, and Pär G Jönsson. Fibers of failure: Classifying errors in predictive processes. *Algorithms*, 13(6):150, 2020.

[8] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730, 2015.

[9] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[10] Marc Coudriau, Abdelkader Lahmadi, and Jerome Francois. Topological analysis and visualisation of network monitoring data: Darknet case study. In *2016 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2016.

[11] Simon DeDeo. Wrong side of the tracks: Big data and protected categories. ithaca, ny: Cornell university library, may 28, 2015, 2015.

[12] Emanuel Derman and Paul Wilmott. The financial modelers' manifesto. *Available at SSRN 1324878*, 2009.

[13] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[14] Kelwin Fernandes, Jaime S Cardoso, and Jessica Fernandes. Transfer learning with partial observability applied to cervical cancer screening. In *Iberian conference on pattern recognition and image analysis*, pages 243–250. Springer, 2017.

[15] Nicholas M Fountain-Jones, Gustavo Machado, Scott Carver, Craig Packer, Mariana Recamonde-Mendoza, and Meggan E Craft. How to make more from exposure data? an integrated machine learning pipeline to predict pathogen exposure. *Journal of Animal Ecology*, 88(10):1447–1461, 2019.

[16] Caleb Geniesse, Olaf Sporns, Giovanni Petri, and Manish Saggar. Generating dynamical neuroimaging spatiotemporal representations (dyneusr) using topological data analysis. *Network Neuroscience*, 3(3):763–778, 2019.

[17] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[18] John C Gower and Gavin JS Ross. Minimum spanning trees and single linkage cluster analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 18(1):54–64, 1969.

[19] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[20] Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*, volume 43. CRC press, 1990.

[21] Karen Sparck Jones. Information retrieval and artificial intelligence. *Artificial Intelligence*, 114(1-2):257–281, 1999.

[22] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[23] Jon Kleinberg. Inherent trade-offs in algorithmic fairness. In *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*, pages 40–40, 2018.

[24] Zachary C Lipton. The doctor just won't accept that! interpretable ml symposium. In *31st conference on neural information processing systems (NIPS 2017), Long Beach, CA, USA. arXiv*, volume 1711, 2017.

[25] Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.

[26] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–631, 2013.

[27] Pek Y Lum, Gurjeet Singh, Alan Lehman, Tigran Ishkanov, Mikael Vejdemo-Johansson, Muthu Alagappan, John Carlsson, and Gunnar Carlsson. Extracting insights from the shape of complex data using topology. *Scientific reports*, 3:1236, 2013.

[28] Microsoft Corporation. *InterpretML*, 2020. https://github.com/interpretml/interpret.

[29] George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

[30] Christoph Molnar. *Interpretable Machine Learning*. 2019. https://christophm.github.io/interpretable-ml-book/ & https://christophm.github.io/interpretable-ml-book/global.html.

[31] Detlef Nauck and Rudolf Kruse. Obtaining interpretable fuzzy classification rules from medical data. *Artificial intelligence in medicine*, 16(2):149–169, 1999.

[32] Monica Nicolau, Arnold J Levine, and Gunnar Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, 108(17):7265–7270, 2011.

[33] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*, 2019.

[34] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 3(3):e10, 2018.

[35] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568, 2008.

[36] Ben Popken. Twitter deleted 200,000 russian troll tweets. read them here. *NBC News*, 14, 2018.

[37] Esther Rolf, Max Simchowitz, Sarah Dean, Lydia T Liu, Daniel Björkegren, Moritz Hardt, and Joshua Blumenstock. Balancing competing objectives with noisy data: Score-based classifiers for welfare-aware machine learning. *arXiv preprint arXiv:2003.06740*, 2020.

[38] Cynthia Rudin and Joanna Radin. Why are we using black box models in ai when we don't need to? a lesson from an explainable ai competition. *Harvard Data Science Review*, 1(2), 11 2019. https://hdsr.mitpress.mit.edu/pub/f9kuryi8.

[39] Karin Sasaki, Dunja Bruder, and Esteban A Hernandez-Vargas. Topological data analysis to model the shape of immune responses during co-infections. *Communications in Nonlinear Science and Numerical Simulation*, 85:105228, 2020.

[40] Nathaniel Saul and Dustin L Arendt. Machine learning explanations with topological data analysis. In *Demo at the Workshop on Visualization for AI explainability (VISxAI)*, 2018.

[41] Andrew Selbst and Julia Powles. "meaningful information" and the right to explanation. In *Conference on Fairness, Accountability and Transparency*, pages 48–48. PMLR, 2018.

[42] Harlan Sexton. The beautiful duality of tda, August 2015. [Online; posted 27-August-2015].

[43] Gurjeet Singh, Facundo Mémoli, and Gunnar E Carlsson. Topological methods for the analysis of high dimensional data sets and 3d object recognition. *SPBG*, 91:100, 2007.

[44] Guillaume Tauzin, Umberto Lupo, Lewis Tunstall, Julian Burella Pérez, Matteo Caorsi, Anibal Medina-Mardones, Alberto Dassatti, and Kathryn Hess. giotto-tda: A topological data analysis toolkit for machine learning and data exploration, 2020.

[45] Ryan Turner. A model explanation system. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2016.

[46] Hendrik Jacob Van Veen, Nathaniel Saul, David Eargle, and Sam W Mangham. Kepler mapper: A flexible python implementation of the mapper algorithm. *Journal of Open Source Software*, 4(42):1315, 2019.

[47] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[48] David H Wolpert. A mathematical theory of generalization. *Complex Systems*, 4(2):151–249, 1990.

[49] I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, 2009.

A Appendix A: Algorithms & Experiments

A.1 Mapper on Charts Algorithmic Details

Algorithm 1 `Mapper` on charts

Result: $\mathcal{H}$. Interface for Interpretable ML
Input: $\mathcal{E}$. 2-$axis$ Chart; X. Data set.
Initialize: Canvas $\mathcal{H}$ from $\mathcal{E}$
for $axis \in \mathcal{E}$ **do**
  $\mathcal{G}(V, E)_{axis} \leftarrow$ `Mapper`(X, lens=$axis$)
  $\mathcal{H}$.`draw`($\mathcal{G}(V, E)_{axis}$)
  for $V_i \in \mathcal{G}(V, E)_{axis}$ **do**
    $\mathcal{H}$.`drawUpdate`(
      Confine V_i to interval on $axis$
      Order V_i in interval by set size
)
    for $U = \{\}, x \in V_i$ **do**
      $u \leftarrow$ `localExplanation`(x)
      U.`update`(u)
    $\mathcal{H}$.`draw`($U, \overline{U}$)

A.2 Experiment reproducibility Checklists

Experiments			
Listing	Cervical Cancer	Loan Data	Propaganda Tweets
Lens horizontal	Age	Row #	Day #
Lens vertical	Risk Score (with uncertainty estimates)	Mean prediction of 100 rows (with stdev prediction of 100 rows)	Log_2 # Tweets per day
# Intervals horizontal	7	14	60
# Intervals vertical	3	None	None
Overlap Percentage	10% / 30%	30%	33%
Clusterer horizontal	Agglomerative Clustering(clusters=2)	Agglomerative Clustering(clusters=2)	DBSCAN
Clusterer vertical	Agglomerative Clustering(clusters=3)	None	None
Color Function	Average Target in Cluster	Stdev of batch of 100 predictions	Distance to mean
Mapper Tuning	Manual visual feedback	None	Manual visual feedback
Mapper Timing	< 1 second	< 10 seconds	< 10 minutes
Model	EBM	EBM	TF-IDF
Model Tuning	Manual XGBoost comparison (See Appendix C)	None	None
Learning Methods	Supervised Semi-Supervised Unsupervised	Supervised Unsupervised	Unsupervised
Environment	python=3.6.8 sklearn=0.20.3 pandas=0.25.1 interpret=0.1.20 kmapper=1.1.5 xgboost=0.81	python=3.6.8 sklearn=0.20.3 pandas=0.25.1 interpret=0.1.20 kmapper=1.1.5 xgboost=0.81	python=3.6.8 sklearn=0.20.3 pandas=0.25.1 kmapper=1.1.5
Hardware	Macbook Pro 2015	Macbook Pro 2015	Macbook Pro 2015

A.3 Mapping details

Note that the authors lack the medical domain expertise necessary to responsibly interpret the resulting visualization for the Cervical Cancer Data Set. The point of that demonstration then is to showcase the interface and its parts, global - to local explanations of samples and features, and the layout of the `Mapper` graphs in relation to the chart.

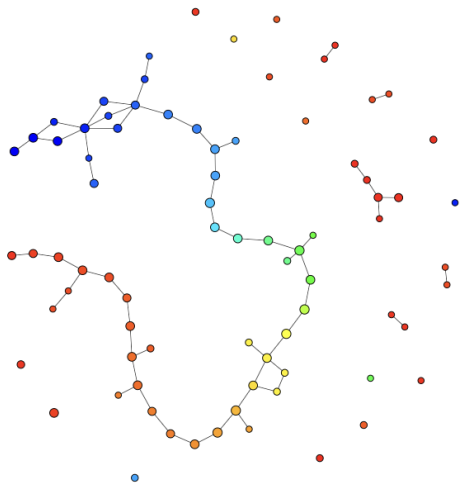
For the Propaganda Tweet data set, we can visualize with KeplerMapper, using data vectorized for words. DBSCAN finds 79 clusters and this results in the following figure A.1 for comparison with our "`Mapper` on charts" methodology:

Figure A.1: KeplerMapper output for Propaganda Tweets experiment.

Cluster Meta [-]

ABOVE AVERAGE

Feature	Mean	STD
trumpforpresident	0.092	2.3x
2016electionin3words	0.05	1.4x
electionday	0.033	1.3x
hillaryforprison2016	0.034	1.3x
icelebratetrumpwith	0.035	1.2x



B Appendix B

B.1 Data Reproducibility Checklists

Data Sets			
Listing	Cervical Cancer	Loan Data	Propaganda Tweets
Number of samples	858	30.000	203.451
Number of samples after processing	849	30.000	120.058
Number of features	32	23	228.130 tokens
Number of features after processing	32	23	3.000
Number of targets	4	1	0
Number of targets after processing	1	1	0
Percentage target	11.7%	22.1%	None
Evaluation splits	Stratified 80% train 20% test	Out-of-time 25.000 train 2.500 test normal 2.500 test with concept drift induced	300 days around the 2016 US election
Sourced	UCI ML repository/ Venezuela hospital/ Patients	UCI ML repository/ Taiwan bank/ Customers	ABCNews/ Twitter/ Alleged Trolls
Link to raw data	Cervical cancer (Risk Factors)	Default of credit card clients	Researcher Tweets.csv

B.2 Cervical Cancer Risk Factors Data Set

This Cervical Cancer Risk Factors data set was introduced in [14]. Some participants declined to answer certain sensitive questions, so there are missing values. It is unexplored if there are problems in the collection of this data set. The data set has 858 rows and 32 features:

```
['Age',
 'Number of sexual partners',
 'First sexual intercourse',
 'Num of pregnancies',
 'Smokes',
 'Smokes (years)',
 'Smokes (packs/year)',
 'Hormonal Contraceptives',
 'Hormonal Contraceptives (years)',
 'IUD',
 'IUD (years)',
 'STDs',
 'STDs (number)',
 'STDs:condylomatosis',
 'STDs:cervical condylomatosis',
 'STDs:vaginal condylomatosis',
 'STDs:vulvo-perineal condylomatosis',
 'STDs:syphilis',
 'STDs:pelvic inflammatory disease',
 'STDs:genital herpes',
 'STDs:molluscum contagiosum',
 'STDs:AIDS',
 'STDs:HIV',
 'STDs:Hepatitis B',
 'STDs:HPV',
 'STDs: Number of diagnosis',
```

```
'STDs: Time since first diagnosis',
'STDs: Time since last diagnosis',
'Dx:Cancer',
'Dx:CIN',
'Dx:HPV',
'Dx']
```

B.2.1 Preprocessing the Cervical Cancer Data

For preprocessing, we crudely impute ? values with 0 values. For binary target creation, we set the target to 1 if any of the 4 target columns (Hinselmann, Schiller, Citology, Biopsy) is 1. Furthermore, for this demonstration, we delete outliers over 50 years of Age.

B.3 Loan Data set

This dataset was collected from a Taiwanese bank in [49] for the purpose of comparing data mining methods. It has 30.000 rows and 23 features:

```
['LIMIT_BAL',
 'SEX',
 'EDUCATION',
 'MARRIAGE',
 'AGE',
 'PAY_0',
 'PAY_2',
 'PAY_3',
 'PAY_4',
 'PAY_5',
 'PAY_6',
 'BILL_AMT1',
 'BILL_AMT2',
 'BILL_AMT3',
 'BILL_AMT4',
 'BILL_AMT5',
 'BILL_AMT6',
 'PAY_AMT1',
 'PAY_AMT2',
 'PAY_AMT3',
 'PAY_AMT4',
 'PAY_AMT5',
 'PAY_AMT6']
```

The average signal "default payment next month" is 22%.

B.3.1 Concept drift induction

To induce concept drift we shuffle the feature columns. We take a holdout set of 5000 of the last rows of the train data and split those in two equal parts. For the first part we do nothing, for the second part, we shuffle 50% of the features in each row.

B.4 Propaganda Tweets

This data set was curated by ABCNews, after actions taken by Twitter as part of their commitment to transparency. There are 393 unique user IDs and 8.065 tweets do not have a user ID attached. There are 453 unique usernames. After removing the tweets without any text we are left with 203.430 tweets. After removing tweets from days with low activity (<30 tweets) we have 200.134 tweets. Of those, there are 172.037 unique tweets. The first tweet was on 10/11/2014 and the last was on 15/08/2017. The mean activity per day is 330 tweets and the median activity per day is 186.

Social media text data is very dynamic and informal. This increases NLP challenges. The data contain mild profanity and relates to sensitive issues. Since:

- this is not an official machine learning data set, created especially for the purpose of data mining,

- there is no way to verify what Twitter has done to detect these alleged trolls,

- it was reported by a popular news media organization classified with an AllSides Media Bias Rating of "Lean Left",

- the data source is connected to state-sponsored propaganda;

it is possible the data contain bias in regards to issues like politics, creation, collection, misdirection, attribution, and curating.

A random sample:

```
"#IAmOnFire Thank you Lord for giving me my abilities, my talents. ALL GLORY IS
    YOURS."

"RT @tlamb775: .@20committee @YouTube Jonasson is anti NATO"

"RT @Lady4Yeshua: #MuslimBan #ExtremeVetting #WomansMarch https://t.co/VtlRUL5llJ"

"RT @jonfavs: This is who Trump is campaigning with today. Called Obama a subhuman
    mongrel and said Hillary should be hanged for "

"RT @BKAdams1984: Rudy Giuliani's conflicts of interest would put Trump in a tough
    bind https://t.co/QMbaMkZoAu via @HuffPostPol"
```

C Modeling and Evaluation Details

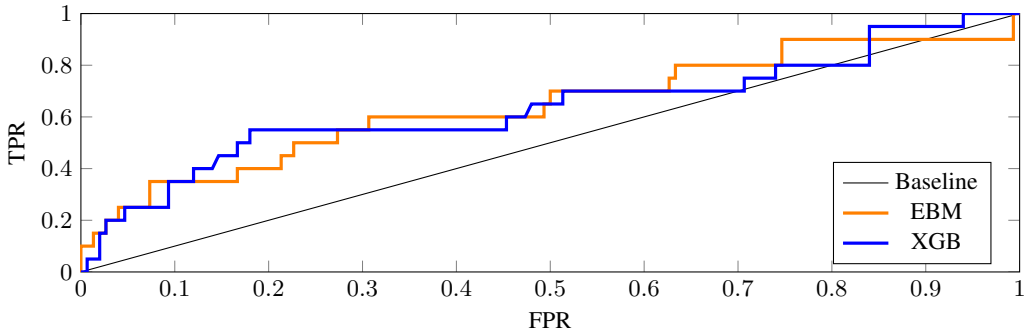
C.1 Modeling reproducibility Checklists

Models			
Listing	Cervical Cancer	Loan Data	Propaganda Tweets
Runtime	EBM: < 1 minute XGB: < 1 second Agg. Clustering: < 10 seconds	EBM: < 1 minute XGB: < 10 seconds Agg. Clustering: < 1 minute	TF-IDF: < 5 minutes DBSCAN: < 10 minutes
Evaluation	Area Under the Curve (AUC) EBM: 0.643 XGB: 0.636	AUC EBM test: 0.798 EBM test drifted: 0.794 EBM test drifted feature switched: 0.686 XGB test: 0.803 XGB test drifted: 0.801 XGB test drifted feature switched: 0.683	Visual Feedback
Parameter Tuning	EBM: manually set 100 estimators, and 2 interactions XGB: None	EBM: None XGB: None	Model - and parameter selection for speed and hardware resources

Same hardware and software environment as Appendix A. Same preprocessing and evaluation splits as Appendix B.

C.2 Cervical Cancer Evaluation

Figure C.1: ROC AUC Curves comparison EBM vs. XGBoost (XGB), Cervical Cancer Risks Experiment



C.3 Loan default

Figure C.2: ROC AUC Curves comparison EBM vs. XGBoost (XGB), Loan Default Concept Drift Experiment