Responsible Machine Learning*

Lecture 6: Responsible Machine Learning Best Practices

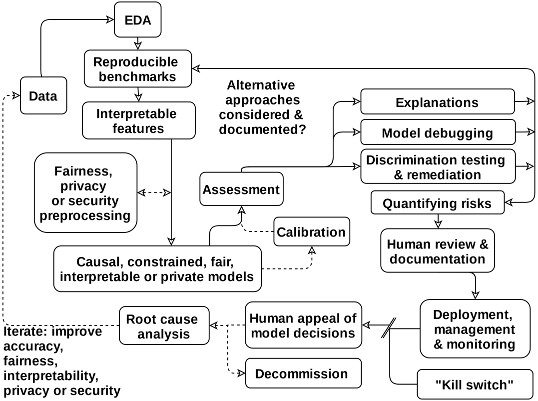
Patrick Hall

The George Washington University

June 24, 2021

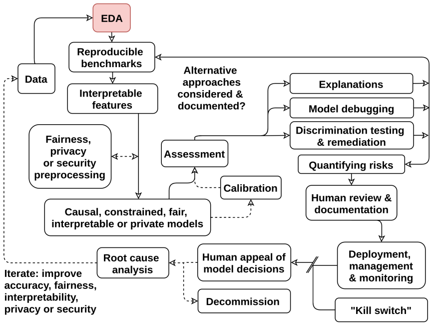
Contents

Responsible ML Blueprint†



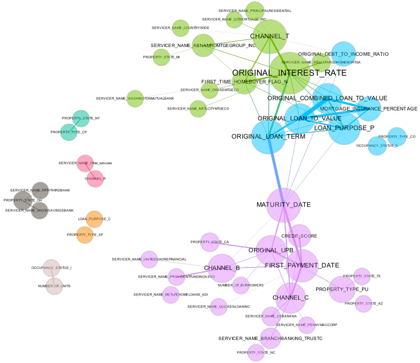
† This blueprint does not address ETL workflows.

EDA and Data Visualization

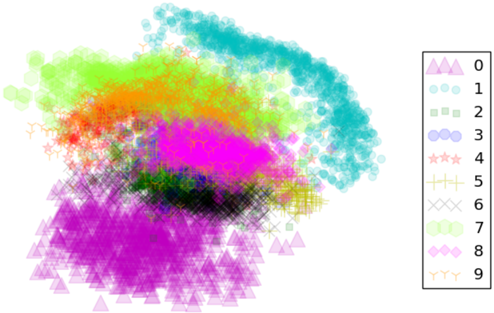


- Know thy data.
- OSS: H2O-3 Aggregator
- References: Visualizing Big Data Outliers through Distributed Aggregation; The Grammar of Graphics

Interlude: My Favorite Visualizations



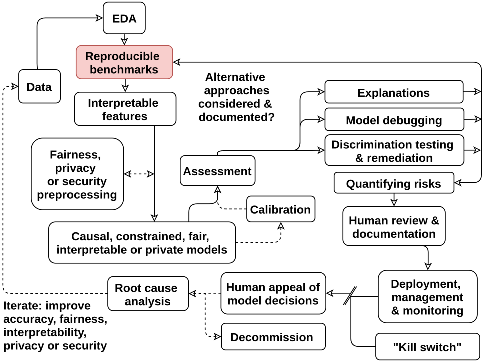
A network graph capturing the Pearson correlation relationships between many *columns* in a lending dataset.



An autoencoder projection of the MNIST data. Projections capture sparsity, clusters, hierarchy, and outliers in *rows* of a dataset.

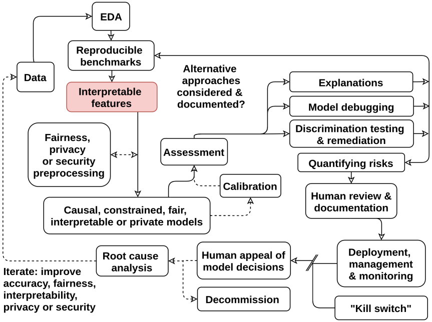
Both of these images capture high-dimensional datasets in just two dimensions.

Establish Benchmarks



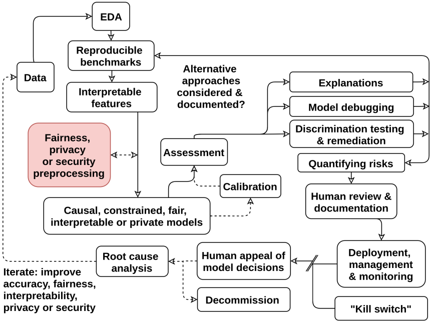
Establishing reproducible benchmarks from which to gauge improvements in accuracy, fairness, interpretability or privacy is crucial for good ("data") science and for compliance.

Manual, Private, Sparse or Straightforward Feature Engineering



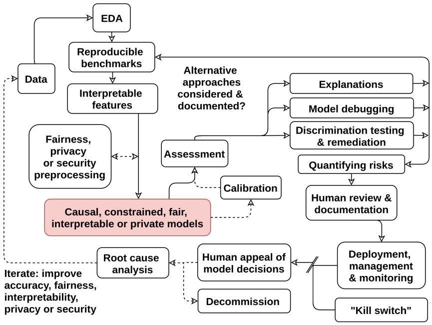
- OSS: elasticnet, Feature Tools
- References: Sparse Principal Component Analysis; Label, Segment, Featurize: A Cross Domain Framework for Prediction Engineering; *t*-Closeness: Privacy Beyond *k*-Anonymity and *l*-diversity

Preprocessing for Fairness, Privacy or Security



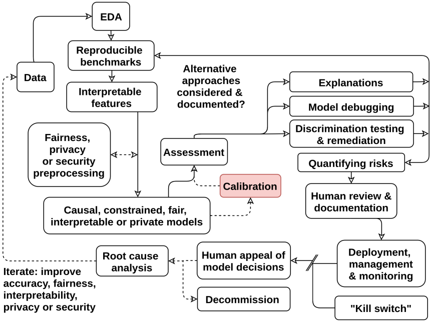
- OSS: IBM AIF360 and diffprivlib
- References: Data Preprocessing Techniques for Classification Without Discrimination; Certifying and Removing Disparate Impact; Optimized Pre-processing for Discrimination Prevention; Privacy-Preserving Data Mining; Differential Privacy and Machine Learning: A Survey and Review

Constrained, Fair, Interpretable, Private or Simple Models



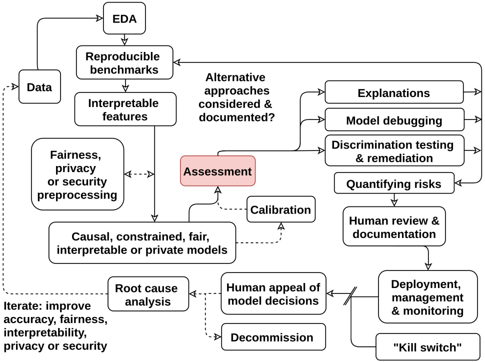
- OSS: Accurate Intelligible Models with Pairwise Interactions (GA2M/EBM); Rudin Group models e.g. Scalable Bayesian Rule Lists (SBRL); Monotonic gradient boosting machines in H2O-3 or XGBoost; pymc3
- References: Scalable Private Learning with PATE; Mitigating Unwanted Biases with Adversarial Learning; Bayesian Networks; Explainable Neural Networks Based on Additive Index Models (XNN)

Prediction Calibration



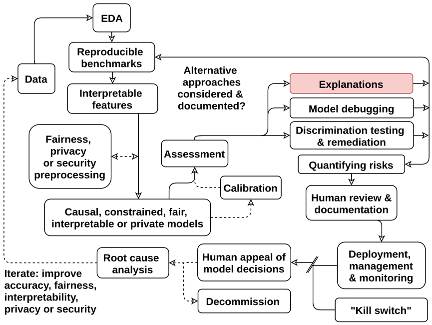
- Just because a number is in $[0, 1]$ does not make it a probability.
- OSS: scikit-learn
- References: Predicting Good Probabilities with Supervised Learning

Traditional Model Assessment and Diagnostics



Residual analysis, Q-Q plots, AUC and lift curves etc. confirm model is accurate and meets assumption criteria.

Post-hoc Explanations

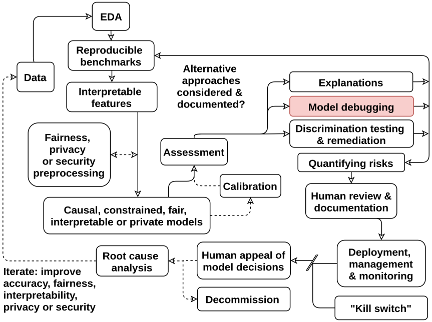


- Explanations enable *understanding* and *appeal* ... *not trust*.
- OSS: alibi, shap
- References: Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GPDR; A Unified Approach to Interpreting Model Predictions; Interpreting Blackbox Models via Model Extraction; Please Stop Explaining Black Box Models for High Stakes Decisions (criticism)

Interlude: The Time–Tested Shapley Value

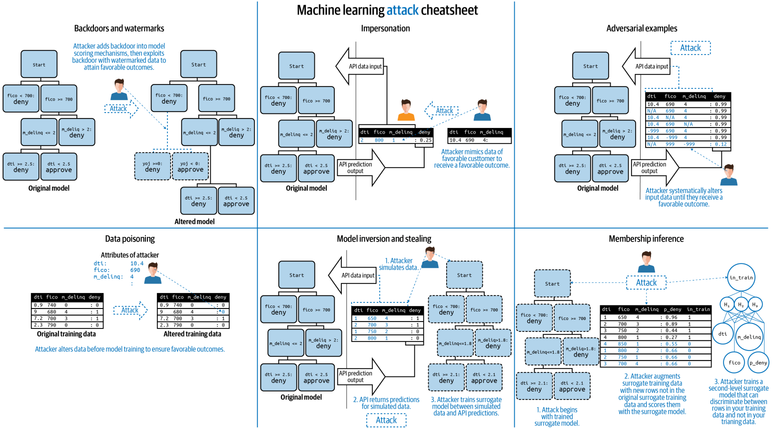
1. **In the beginning**: A Value for N-Person Games, 1953
2. **Nobel-worthy contributions**: The Shapley Value: Essays in Honor of Lloyd S. Shapley, 1988
3. **Shapley regression**: Analysis of Regression in Game Theory Approach, 2001
4. **First reference in ML?** Fair Attribution of Functional Contribution in Artificial and Biological Networks, 2004
5. **Into the ML research mainstream, i.e. JMLR**: An Efficient Explanation of Individual Classifications Using Game Theory, 2010
6. **Into the real-world data mining workflow ... *finally***: Consistent Individualized Feature Attribution for Tree Ensembles, 2017
7. **Unification**: A Unified Approach to Interpreting Model Predictions, 2017

Model Debugging for Accuracy, Privacy or Security



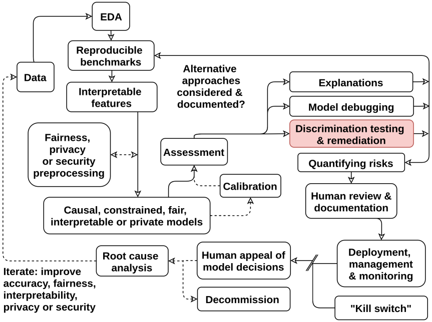
- Eliminating errors in model predictions by testing: adversarial examples, explanation of residuals, random attacks and "what-if" analysis.
- OSS: cleverhans, pdpbox, what-if tool, robustness
- References: Modeltracker: Redesigning Performance Analysis Tools for Machine Learning; A Marauder's Map of Security and Privacy in Machine Learning: An overview of current and future research directions for making machine learning secure and private; The Security of Machine Learning

Machine Learning Attacks‡



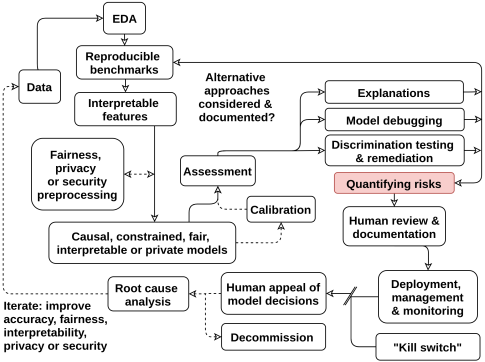
‡ See https://github.com/jphall663/secure_ML_ideas for full size image and more information.

Post-hoc Disparate Impact Assessment and Remediation



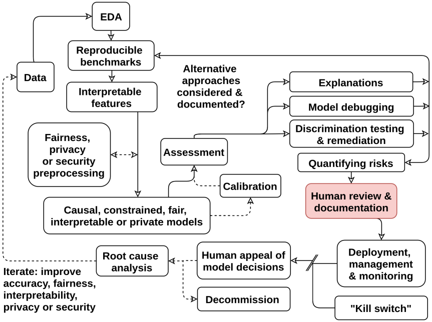
- Social bias testing should include group fairness tests and should attempt to consider individual fairness.
- OSS: aequitas, IBM AIF360, themis
- References: Fairness Through Awareness; Decision Theory for Discrimination-aware Classification; Equality of Opportunity in Supervised Learning; Certifying and Removing Disparate Impact

Quantify and Plan for Risk



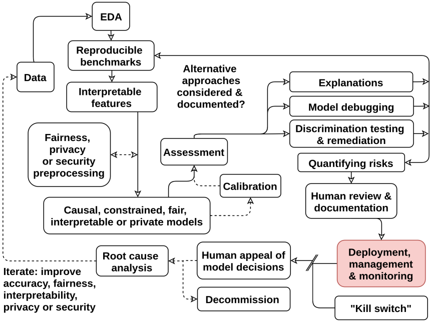
Your model will be wrong. Stake-holders need to understand and be prepared for the human and financial costs of these wrong decisions.

Human Review and Documentation



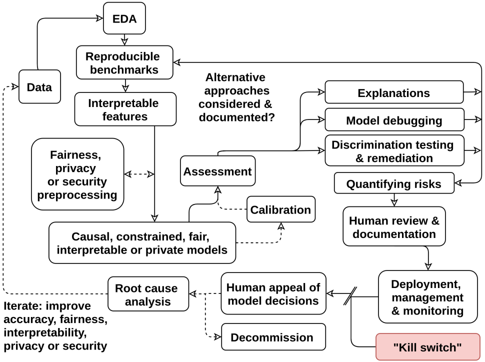
- Reference: Model Cards for Model Reporting
- Documentation of considered alternative approaches typically necessary for compliance.

Deployment, Management and Monitoring



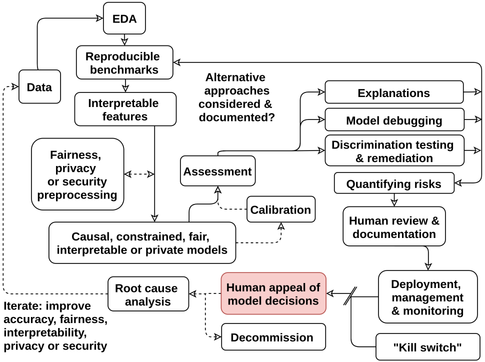
- Monitor models for accuracy, disparate impact, privacy violations or security vulnerabilities in real-time; track model and data lineage.
- OSS: DVC, gigantum, KubeFlow, mlflow, modeldb, TensorFlow ML Metadata, TensorFlow TFX, awesome-machine-learning-ops metalist
- Reference: Model DB: A System for Machine Learning Model Management

Kill Switches



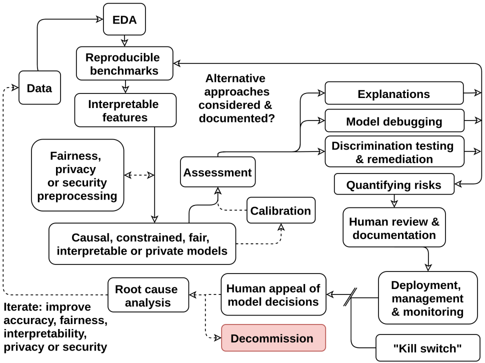
Being able to quickly turn off a misbehaving ML system is crucially important. This requires technical and organizational considerations. E.g., how much revenue is lost each minute a model is disabled?

Human Appeal



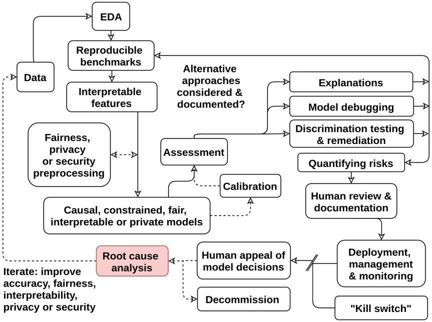
Very important, may require custom implementation for each deployment environment? Related problems exist *today*.

Decommission Model



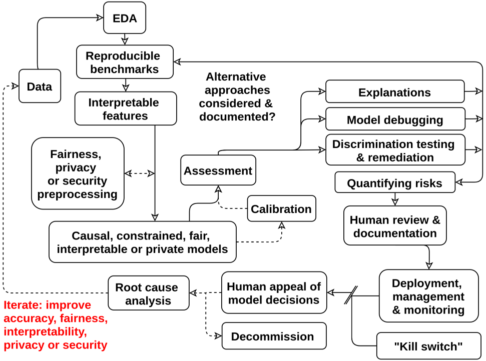
When a model becomes absolutely or relatively inaccurate, unfair, or insecure it must be taken out of service, but saved in an executable and reproducible manner.

Causality?



- Root cause analysis: can root causes be identified, verified? Formalized into model architecture?
- OSS: dowhy, pymc3
- References: The Book of Why: the New Science of Cause and Effect; Probabilistic Programming in Python using PyMC3

Iterate: Use Gained Knowledge to Improve Accuracy, Fairness, Interpretability, Privacy or Security



Improvements, KPIs should not be restricted to accuracy alone.

Process Solutions

- **Bug Bounties**: Offer rewards to the broader community to find all kinds of problems (discrimination, opacity, vulnerabilities, privacy harms, etc.) in your organization's public-facing ML systems.
- **Data and AI Principles**: Devise central tenants for how your organization will handle ethical, political, and legal issues related to data and ML.
- **Diversity of Experience**: Ensure data and ML teams are staffed with individuals that can share different demographic, technical, and professional perspectives.

Process Solutions

- **"Dog-fooding"**: If possible, test your ML system on yourself or internally at your organization. Don't feel comfortable using it on yourself? Maybe you shouldn't release it.
- **Documentation**: Documentation ends up being the primary physical implementation of many risk controls.
- **Domain Expertise**: Success in ML almost always requires input from humans with deep understanding of the problem domain.
- **Effective Challenge and Human Review**: Nearly all aspects of ML workflows should involve challenges and questioning from group members. This can be in the form of human interrogation of ML-related processes or in the form of challenger models.
- **Executive Oversight**: An empowered executive with a staff and budget can exert a strong influence over organizational use of ML.

Process Solutions

- **Incident Response Plans**: Complex ML systems *will* fail. Being prepared for failures or attacks can be the difference between a major incident and a minor disruption.
- **Incentives**: Model builders, testers, auditors, and executives all have different roles to play in the implementation of responsible ML and should be incentivized to play the correct role.
- **Legal Privilege**: Consider use of privilege to minimize risk when dealing with ML-related legal and compliance issues.
- **Model Risk Management**: The established practice of model risk management can be expanded outside of financial services.
- **Red-teaming**: Establish a group or hire third-parties to act as adversaries and find problems (discrimination, opacity, vulnerabilities, privacy harms, etc.) in your organization's public-facing ML systems.

Acknowledgments

Thanks to Lisa Song for her continued assistance in developing these course materials.

References

Agrawal, Rakesh and Ramakrishnan Srikant (2000). "Privacy-Preserving Data Mining." In: *ACM Sigmod Record*. Vol. 29. 2. URL: http://alme1.almaden.ibm.com/cs/projects/iis/hdb/Publications/papers/sigmod00_privacy.pdf. ACM, pp. 439–450.

Amershi, Saleema et al. (2015). "Modeltracker: Redesigning Performance Analysis Tools for Machine Learning." In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. URL: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/amershi.CHI2015.ModelTracker.pdf. ACM, pp. 337–346.

Barreno, Marco et al. (2010). "The Security of Machine Learning." In: *Machine Learning* 81.2. URL: https://people.eecs.berkeley.edu/~adj/publications/paper-files/SecML-MLJ2010.pdf, pp. 121–148.

Bastani, Osbert, Carolyn Kim, and Hamsa Bastani (2017). "Interpreting Blackbox Models via Model Extraction." In: *arXiv preprint arXiv:1705.08504*. URL: https://arxiv.org/pdf/1705.08504.pdf.

Calmon, Flavio et al. (2017). "Optimized Pre-processing for Discrimination Prevention." In: *Advances in Neural Information Processing Systems*. URL: http://papers.nips.cc/paper/6988-optimized-pre-processing-for-discrimination-prevention.pdf, pp. 3992–4001.

References

Dwork, Cynthia et al. (2012). "Fairness Through Awareness." In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. URL: https://arxiv.org/pdf/1104.3913.pdf. ACM, pp. 214–226.

Feldman, Michael et al. (2015). "Certifying and Removing Disparate Impact." In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. URL: https://arxiv.org/pdf/1412.3756.pdf. ACM, pp. 259–268.

Hardt, Moritz, Eric Price, Nati Srebro, et al. (2016). "Equality of Opportunity in Supervised Learning." In: *Advances in neural information processing systems*. URL: http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf, pp. 3315–3323.

Ji, Zhanglong, Zachary C. Lipton, and Charles Elkan (2014). "Differential Privacy and Machine Learning: A Survey and Review." In: *arXiv preprint arXiv:1412.7584*. URL: https://arxiv.org/pdf/1412.7584.pdf.

Kamiran, Faisal and Toon Calders (2012). "Data Preprocessing Techniques for Classification Without Discrimination." In: *Knowledge and Information Systems* 33.1. URL: https://link.springer.com/content/pdf/10.1007/s10115-011-0463-8.pdf, pp. 1–33.

References

Kamiran, Faisal, Asim Karim, and Xiangliang Zhang (2012). "Decision Theory for Discrimination-aware Classification." In: *2012 IEEE 12th International Conference on Data Mining*. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.722.3030&rep=rep1&type=pdf. IEEE, pp. 924–929.

Kanter, James Max, Owen Gillespie, and Kalyan Veeramachaneni (2016). "Label, Segment, Featurize: A Cross Domain Framework for Prediction Engineering." In: *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*. URL: http://www.jmaxkanter.com/static/papers/DSAA_LSF_2016.pdf. IEEE, pp. 430–439.

Keinan, Alon et al. (2004). "Fair Attribution of Functional Contribution in Artificial and Biological Networks." In: *Neural Computation* 16.9. URL: https://bit.ly/35kHrwK, pp. 1887–1915.

Kononenko, Igor et al. (2010). "An Efficient Explanation of Individual Classifications Using Game Theory." In: *Journal of Machine Learning Research* 11.Jan. URL: http://www.jmlr.org/papers/volume11/strumbelj10a/strumbelj10a.pdf, pp. 1–18.

Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian (2007). "t-Closeness: Privacy Beyond k-Anonymity and l-diversity." In: *2007 IEEE 23rd International Conference on Data Engineering*. URL:http://www.utdallas.edu/~mxk055100/courses/privacy08f_files/tcloseness.pdf. IEEE, pp. 106–115.

References

Lipovetsky, Stan and Michael Conklin (2001). "Analysis of Regression in Game Theory Approach." In: *Applied Stochastic Models in Business and Industry* 17.4, pp. 319–330.

Lou, Yin et al. (2013). "Accurate Intelligible Models with Pairwise Interactions." In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.352.7682&rep=rep1&type=pdf. ACM, pp. 623–631.

Lundberg, Scott M., Gabriel G. Erion, and Su-In Lee (2017). "Consistent Individualized Feature Attribution for Tree Ensembles." In: *Proceedings of the 2017 ICML Workshop on Human Interpretability in Machine Learning (WHI 2017)*. Ed. by Been Kim et al. URL: https://openreview.net/pdf?id=ByTKSo-m-. ICML WHI 2017, pp. 15–21.

Lundberg, Scott M and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions." In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf. Curran Associates, Inc., pp. 4765–4774.

Mitchell, Margaret et al. (2019). "Model Cards for Model Reporting." In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. URL: https://arxiv.org/pdf/1810.03993.pdf. ACM, pp. 220–229.

References

Niculescu-Mizil, Alexandru and Rich Caruana (2005). "Predicting Good Probabilities with Supervised Learning." In: *Proceedings of the 22nd international conference on Machine learning*. URL: https://bit.ly/3434dZG. ACM, pp. 625–632.

Papernot, Nicolas (2018). "A Marauder's Map of Security and Privacy in Machine Learning: An overview of current and future research directions for making machine learning secure and private." In: *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*. URL: https://arxiv.org/pdf/1811.01134.pdf. ACM.

Papernot, Nicolas et al. (2018). "Scalable Private Learning with PATE." In: *arXiv preprint arXiv:1802.08908*. URL: https://arxiv.org/pdf/1802.08908.pdf.

Pearl, Judea (2011). "Bayesian Networks." In: URL: https://escholarship.org/uc/item/53n4f34m.

Pearl, Judea and Dana Mackenzie (2018). *The Book of Why: the New Science of Cause and Effect*. URL: http://cdar.berkeley.edu/wp-content/uploads/2017/04/Lisa-Goldberg-reviews-The-Book-of-Why.pdf. Basic Books.

Rudin, Cynthia (2018). "Please Stop Explaining Black Box Models for High Stakes Decisions." In: *arXiv preprint arXiv:1811.10154*. URL: https://arxiv.org/pdf/1811.10154.pdf.

Salvatier, John, Thomas V Wiecki, and Christopher Fonnesbeck (2016). "Probabilistic Programming in Python using PyMC3." In: *PeerJ Computer Science* 2. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.800.6956&rep=rep1&type=pdf, e55.

References

Shapley, Lloyd S (1953). "A Value for N-Person Games." In: *Contributions to the Theory of Games* 2.28. URL: http://www.library.fa.ru/files/Roth2.pdf#page=39, pp. 307–317.

Shapley, Lloyd S, Alvin E Roth, et al. (1988). *The Shapley Value: Essays in Honor of Lloyd S. Shapley*. URL: http://www.library.fa.ru/files/Roth2.pdf. Cambridge University Press.

Vartak, Manasi et al. (2016). "Model DB: A System for Machine Learning Model Management." In: *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*. URL: https://www-cs.stanford.edu/~matei/papers/2016/hilda_modeldb.pdf. ACM, p. 14.

Vaughan, Joel et al. (2018). "Explainable Neural Networks Based on Additive Index Models." In: *arXiv preprint arXiv:1806.01933*. URL: https://arxiv.org/pdf/1806.01933.pdf.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell (2017). "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GPDR." In: *Harv. JL & Tech.* 31, p. 841. URL: URL:\url{https://arxiv.org/pdf/1711.00399.pdf}.

Wilkinson, Leland (2006). *The Grammar of Graphics*.

— (2018). "Visualizing Big Data Outliers through Distributed Aggregation." In: *IEEE Transactions on Visualization & Computer Graphics*. URL: https://www.cs.uic.edu/~wilkinson/Publications/outliers.pdf.

References

Yang, Hongyu, Cynthia Rudin, and Margo Seltzer (2017). "Scalable Bayesian Rule Lists." In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. URL: https://arxiv.org/pdf/1602.08610.pdf.

Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell (2018). "Mitigating Unwanted Biases with Adversarial Learning." In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. URL: https://bit.ly/2KAV1nG. ACM, pp. 335–340.

Zou, Hui, Trevor Hastie, and Robert Tibshirani (2006). "Sparse Principal Component Analysis." In: *Journal of computational and graphical statistics* 15.2. URL: https://www.cc.gatech.edu/home/isbell/classes/reading/papers/sparsepc.pdf, pp. 265–286.