

# Beyond Reason Codes

## A Blueprint for Human-Centered, Low-Risk AutoML

H2O.ai Machine Learning Interpretability Team

H<sub>2</sub>O.ai

February 1, 2019

# Contents

Blueprint

EDA

Training

Post-Hoc Analysis

Human Review

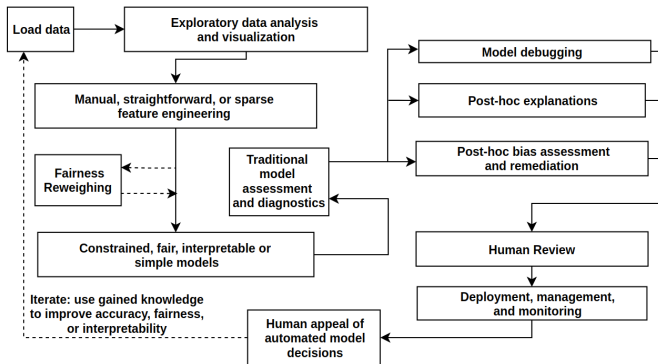
Deployment

Human Appeal

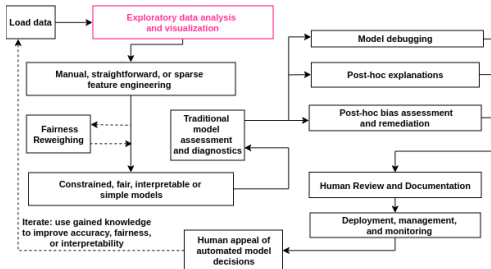
Iterate

Open Questions

# Blueprint

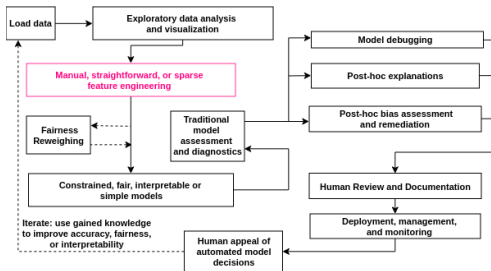


## EDA and Data Visualization



- **Automation** implemented in Driverless AI as AutoViz.
- OSS: **H2O-3 Aggregator**
- References: Visualizing Big Data Outliers through Distributed Aggregation; The Grammar of Graphics

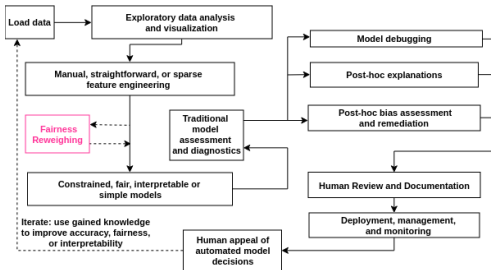
# Manual, Straightforward, or Sparse Feature Engineering



- **Automation** implemented in Driverless AI as high-interpretability transformers.
- OSS: **Pandas Profiler**, **Feature Tools**
- References: Deep Feature Synthesis: Towards Automating Data Science Endeavors; Label, Segment, Featurize: A Cross Domain Framework for Prediction Engineering

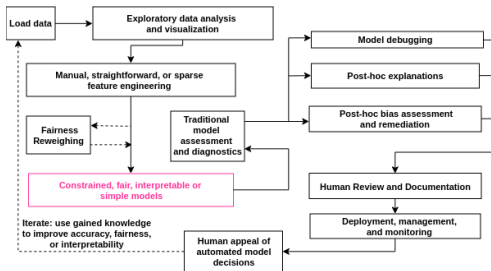


# Fairness Reweighing



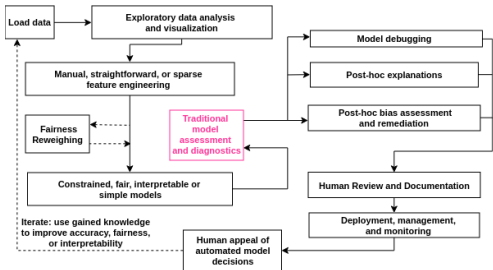
- OSS: IBM **AI360**
- References: Three Naïve Bayes Approaches for Discrimination-free Classification; Data Preprocessing Techniques for Classification Without Discrimination; Certifying and Removing Disparate Impact; Optimized Pre-processing for Discrimination Prevention
- Roadmap items for MLI-2.

# Constrained, Fair, Interpretable or Simple Models



- **Automation** implemented in Driverless AI as GLM, RuleFit, Monotonic GBM.
- References: Locally Interpretable Models and Effects Based on Supervised Partitioning (LIME-SUP); Explainable Neural Networks Based on Additive Index Models (XNN); Scalable Bayesian Rule Lists (SBRL)
- LIME-SUP, SBRL, XNN are roadmap items for MLI-2.

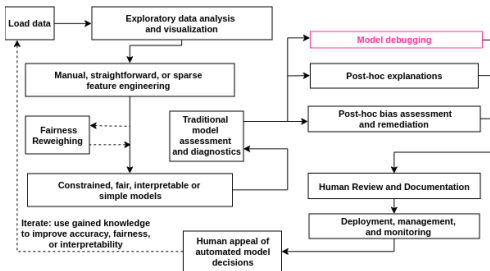
## Traditional Model Assessment and Diagnostics



- Residual analysis, Q-Q plots, AUC and lift curves confirm model is accurate and meets assumption criteria.
- Implemented as model diagnostics in Driverless AI.
- **Residual analysis is roadmap item for model diagnostics in Driverless AI.**

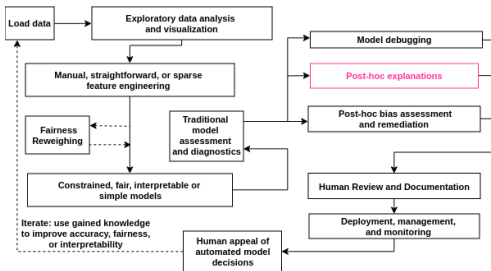


# Model Debugging



- Understanding and eliminating errors in model predictions by model testing: adversarial examples, "what-if" analysis, random attacks, explanation of residuals.
- OSS: **cleverhans**, **pdpbox**, **what-if tool**
- **Adversarial examples**, "what-if" analysis, explanation of residuals, measures of epistemic uncertainty are implemented and roadmap items in MLI-2.

## Post-hoc Explanations



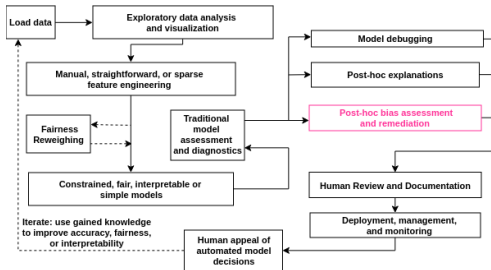
- LIME and Tree SHAP implemented Driverless AI.
- OSS: [lime](#), [shap](#)
- References: Why Should I Trust You?: Explaining the Predictions of Any Classifier; A Unified Approach to Interpreting Model Predictions; Please Stop Explaining Black Box Models for High Stakes Decisions (criticism)
- Tree SHAP is roadmap item for H2O-3; **Explanations for unstructured data are roadmap for MLI-2.**



## Interlude: The Time-Tested Shapley Value

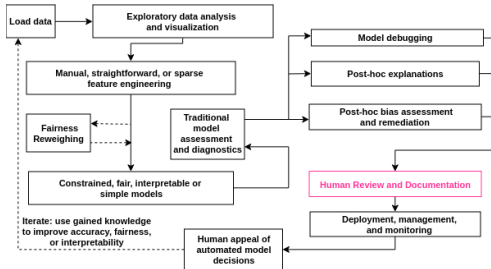
1. **In the beginning:** A Value for N-Person Games, 1953
2. **Nobel-worthy contributions:** The Shapley Value: Essays in Honor of Lloyd S. Shapley, 1988
3. **Shapley regression:** Analysis of Regression in Game Theory Approach, 2001
4. **First reference in ML?** Fair Attribution of Functional Contribution 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

## Post-hoc Disparate Impact Assessment and Remediation



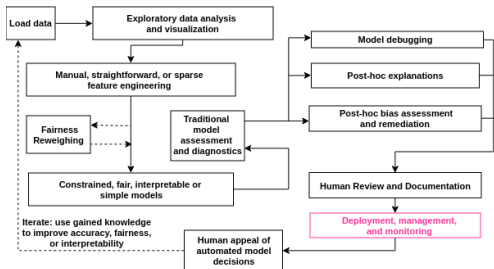
- Disparate impact analysis can be performed manually using Driverless AI or H2O-3.
- OSS: [aequitas](#), IBM [AI360](#), [themis](#)
- References: Equality of Opportunity in Supervised Learning; Certifying and Removing Disparate Impact
- **Disparate impact analysis and remediation** are roadmap items for MLI-2.

## Human Review and Documentation



- Implemented as AutoDoc in Driverless AI.
- **Various interpretability and fairness roadmap items to be added to AutoDoc.**

## Deployment, Management, and Monitoring

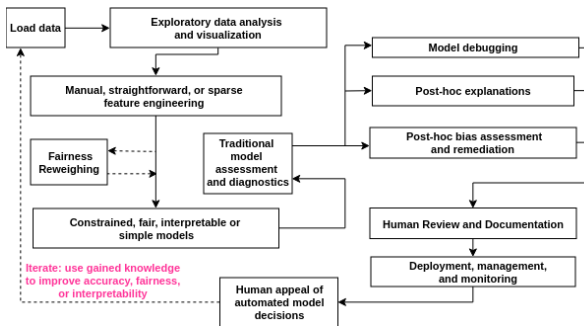


- Monitor models for accuracy and fairness in real-time, track model and data lineage.
- OSS: **mlflow**, **modeldb**
- Reference: Model DB: A System for Machine Learning Model Management
- **Broader roadmap item for H2O.ai.**



Very important, may require custom implementation for each deployment environment?

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



Improvements, KPIs should not be restricted to accuracy alone.



## Open Conceptual Questions

- How much automation is appropriate, 100%?
- How to automate learning by iteration, reinforcement learning?
- How to implement human appeals, is it productizable?

## References

### **Driverless AI API Interpretability Technique Examples:**

<https://github.com/h2oai/driverlessai-tutorials>

### **In-Depth Open Source Interpretability Technique Examples:**

[https://github.com/jphall663/interpretable\\_machine\\_learning\\_with\\_python](https://github.com/jphall663/interpretable_machine_learning_with_python)

### **"Awesome" Machine Learning Interpretability Resource List:**

<https://github.com/jphall663/awesome-machine-learning-interpretability>

## References

- Calders, Toon and Sicco Verwer (2010). “Three Naïve Bayes Approaches for Discrimination-free Classification.” In: *Data Mining and Knowledge Discovery* 21.2. URL: <https://link.springer.com/content/pdf/10.1007/s10618-010-0190-x.pdf>, pp. 277–292.
- 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.
- 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.
- Hu, Linwei et al. (2018). “Locally Interpretable Models and Effects Based on Supervised Partitioning (LIME-SUP).” In: *arXiv preprint arXiv:1806.00663*. URL: <https://arxiv.org/ftp/arxiv/papers/1806/1806.00663.pdf>.

## References

- 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.
- 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](http://www.jmaxkanter.com/static/papers/DSAA_LSF_2016.pdf). IEEE, pp. 430–439.
- Kanter, James Max and Kalyan Veeramachaneni (2015). “Deep Feature Synthesis: Towards Automating Data Science Endeavors.” In: *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*. URL: [https://groups.csail.mit.edu/EVO-DesignOpt/groupWebSite/uploads/Site/DSAA\\_DSM\\_2015.pdf](https://groups.csail.mit.edu/EVO-DesignOpt/groupWebSite/uploads/Site/DSAA_DSM_2015.pdf). IEEE, pp. 1–10.
- Keinan, Alon et al. (2004). “Fair Attribution of Functional Contribution in Artificial and Biological Networks.” In: *Neural Computation* 16.9. URL: [https://www.researchgate.net/profile/Isaac\\_Meilijson/publication/2474580\\_Fair\\_Attribution\\_of\\_Functional\\_Contribution\\_in\\_Artificial\\_and\\_Biological\\_Networks/links/09e415146df8289373000000/Fair-Attribution-of-Functional-Contribution-in-Artificial-and-Biological-Networks.pdf](https://www.researchgate.net/profile/Isaac_Meilijson/publication/2474580_Fair_Attribution_of_Functional_Contribution_in_Artificial_and_Biological_Networks/links/09e415146df8289373000000/Fair-Attribution-of-Functional-Contribution-in-Artificial-and-Biological-Networks.pdf), pp. 1887–1915.

## References

- 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.
- 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.
- 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.
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). "Why Should I Trust You?: Explaining the Predictions of Any Classifier." In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. URL: <http://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf>. ACM, pp. 1135–1144.

## References

- 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>.
- Shapley, Lloyd S (1953). “A Value for N-Person Games.” In: *Contributions to the Theory of Games* 2.28. URL: <http://www.library.f.a.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.f.a.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](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>.
- Wilkinson, Leland (2006). *The Grammar of Graphics*.
- (2018). “Visualizing Big Data Outliers through Distributed Aggregation.” In: *IEEE Transactions on Visualization & Computer Graphics* 1. URL: <https://www.cs.uic.edu/~wilkinson/Publications/outliers.pdf>, pp. 1–1.
- 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>.