

# Beyond Reason Codes

A Blueprint for Human-Centered, Low-Risk Machine Learning

H2O.ai Machine Learning Interpretability Team

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

January 30, 2019

# Contents

[Blueprint](#)

[EDA](#)

[Training](#)

[Post-Hoc Analysis](#)

[Human Review](#)

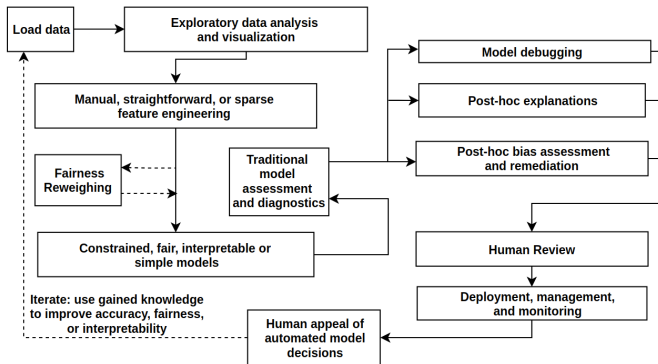
[Deployment](#)

[Human Appeal](#)

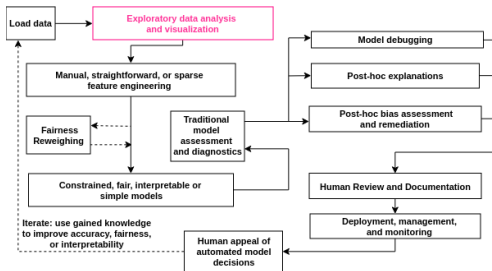
[Iterate](#)

[Open Questions](#)

# Blueprint

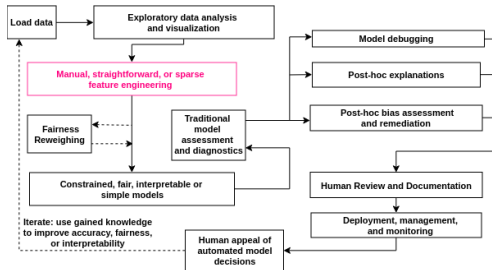


# EDA and Data Visualization



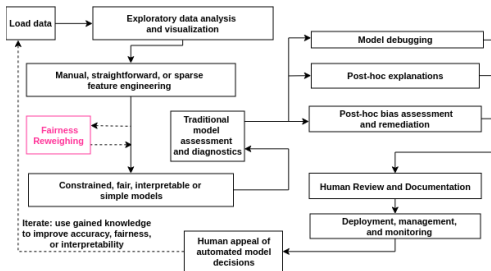
- Implemented in Driverless AI as AutoViz
- OSS: **ggplot**, **seaborn**, etc.
- Reference: *The Grammar of Graphics*, Wilkinson, 2006

# Manual, Straightforward, or Sparse Feature Engineering



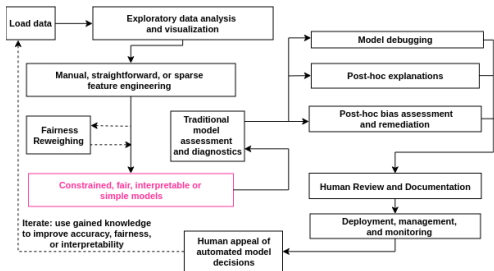
- Implemented in Driverless AI as high-interpretability transformers: frequency, interactions, (monotonic) weight-of-evidence, lags, basics and some Easter eggs in H2O-3
- Decades of custom coding in Hadoop, Python, R, SAS, Spark, SQL, etc.
- Open benchmark of common tools

# Fairness Reweighting



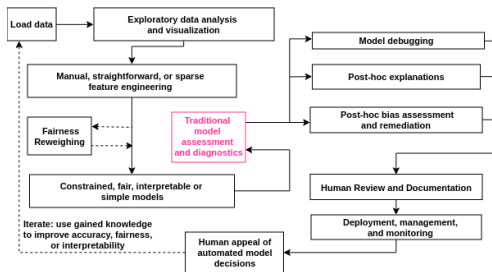
- Newer techniques for reweighing data prior to training to remove disparate impact analysis.
- OSS: IBM **AI360**
- References: Calders and Verwer, **2010**, Kamiran and Calders, **2012**, Feldman et al., **2015**, Calmon et al., **2017**
- **Roamap** items for MLI-2

# Constrained, Fair, Interpretable or Simple Models



- For best transparency use constrained, simple, or directly interpretable models from the beginning
- Implemented in Driverless AI as GLM, RuleFit, Monotonic GBM, in H2O-3 as GLM, monotonic GBM
- Decision tree, scalable Bayesian rulelist, XNN are roadmap items for MLI-2

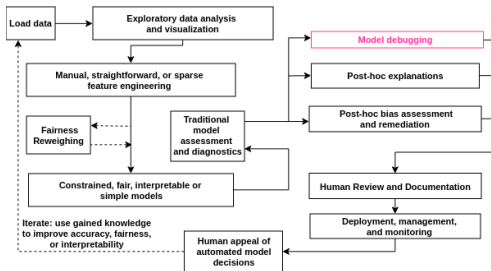
# Traditional Model Assessment and Diagnostics



- Confirms 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

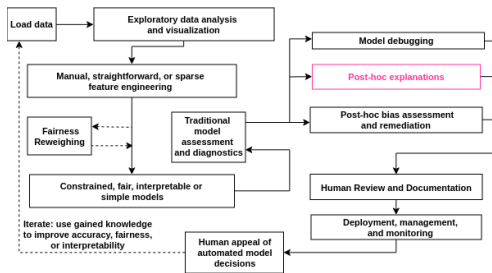


# Model Debugging



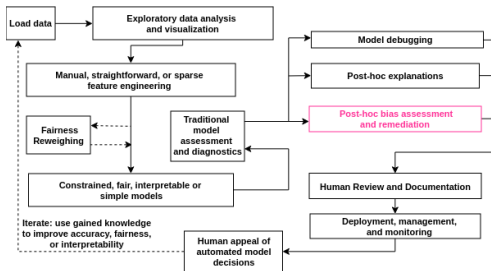
- Newer techniques concerned with understanding and eliminating errors in model predictions; also model testing: "what-if" analysis, random attacks; focus on enhancing *trust*
- "what-if" analysis, explanation of residuals, measures of epistemic uncertainty are roadmap items for MLI-2

# Post-hoc Explanations



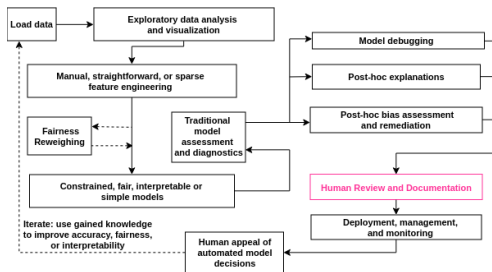
- Explanations enhance *understanding*
- Global feature importance, surrogate decision tree, LIME, LOCO, treeinterpreter and Shapley local feature importance, partial dependence and ICE implemented in current MLI, Friedman's H-statistic implemented in MLI-2
- Shapley is roadmap item for H2O-3; Basic term weights, ALE plots, decision boundary plots are roadmap items for MLI-2

# Post-hoc Disparate Impact Assessment and Remediation



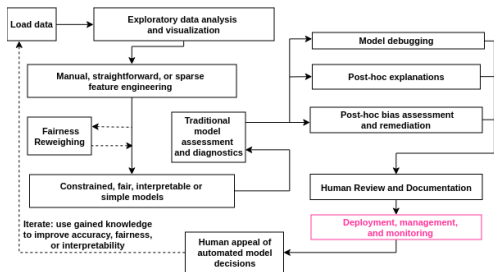
- Disparate Impact Analysis available through code APIs in Driverless AI and H2O-3
- Newer techniques can remove certain types of disparate impact
- Disparate impact remediation is a roadmap item for MLI-2

# Human Review and Documentation



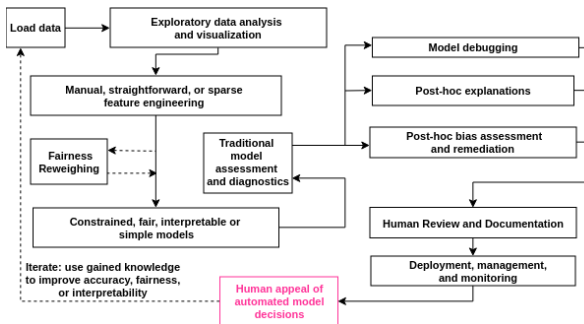
- Implemented as AutoDoc in Driverless AI
- Results from various roadmap items to be added to AutoDoc as appropriate

# Deployment, Management, and Monitoring



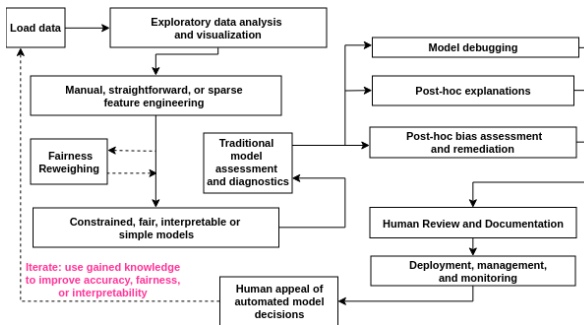
- Monitor models for accuracy and fairness in real-time
- Broader roadmap item for H2O as a company

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



Very important, but probably requires custom implementation for each deployment

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



Improvements, KPIs should not be restricted to accuracy alone

# Open Questions

- What is the role for automation?
- How to implement human appeals, is it productizable?



## References

- Calders, Toon and Sicco Verwer (2010). “Three Naive Bayes Approaches for Discrimination-free Classification.” In: *Data Mining and Knowledge Discovery* 21.2, pp. 277–292.
- Calmon, Flavio et al. (2017). “Optimized pre-processing for discrimination prevention.” In: *Advances in Neural Information Processing Systems*, 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*. ACM, pp. 259–268.
- Kamiran, Faisal and Toon Calders (2012). “Data Preprocessing Techniques for Classification Without Discrimination.” In: *Knowledge and Information Systems* 33.1, pp. 1–33.
- Wilkinson, Leland (2006). *The Grammar of Graphics*. Springer Science & Business Media.