Responsible Machine Learning*

A Blueprint for Human Trust and Understanding in Real-World Machine Learning Systems

Navdeep Gill

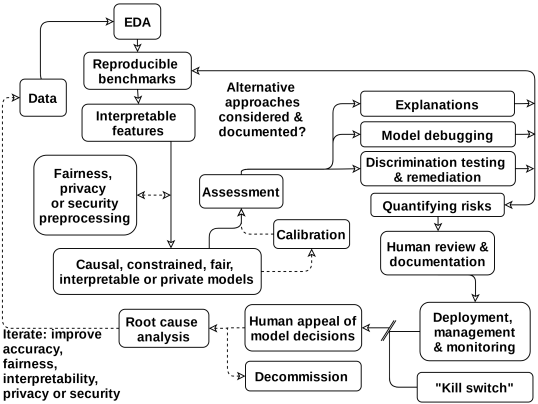
H2O.ai

March 11, 2022

*This material is shared under a CC By 4.0 license which allows for editing and redistribution, even for commercial purposes. However, any derivative work should attribute the author.

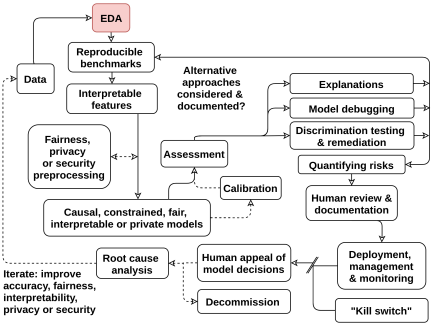
Contents

Responsible ML Blueprint†



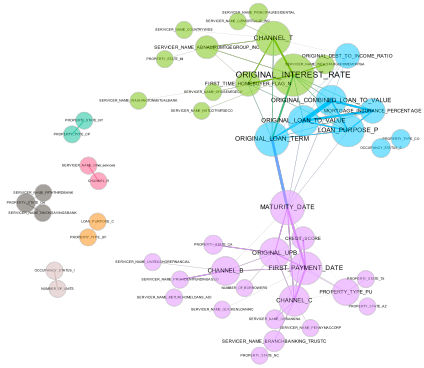
† This blueprint does not address ETL workflows.

EDA and Data Visualization

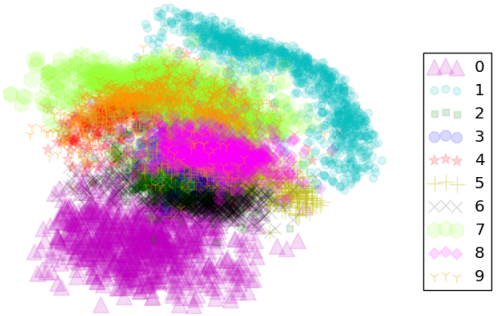


- Know thy data.
- OSS: H2O-3 Aggregator
- References: **wilkinson2018visualizing**; **wilkinson2006grammar**

Interlude: My Favorite Visualizations



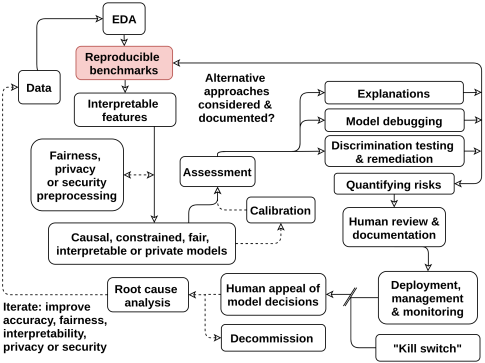
A network graph capturing the Pearson correlation relationships between many *columns* in a lending dataset.



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

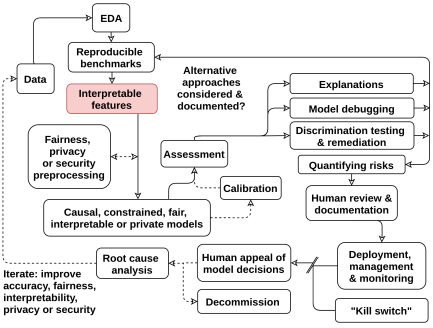
Both of these images capture high-dimensional datasets in just two dimensions.

Establish Benchmarks



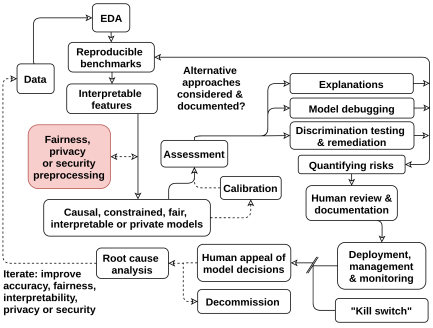
Establishing reproducible benchmarks from which to gauge improvements in accuracy, fairness, interpretability or privacy is crucial for good ("data") science and for compliance.

Manual, Private, Sparse or Straightforward Feature Engineering



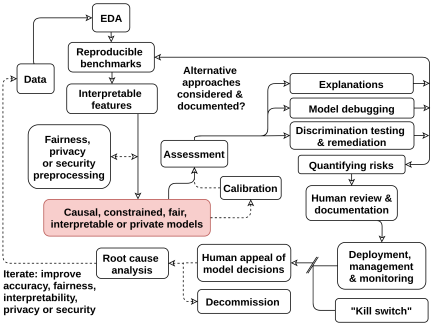
- OSS: elasticnet, Feature Tools
- References: **zou2006sparse**; **kanter2016label**; **t_closeness**

Preprocessing for Fairness, Privacy or Security



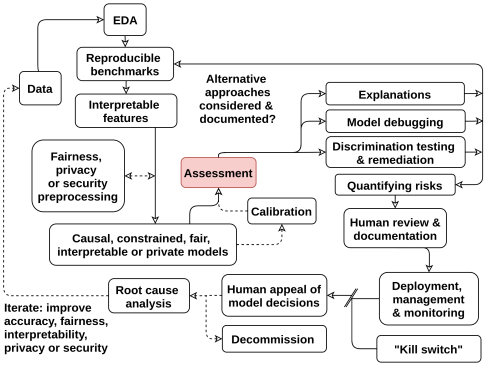
- OSS: IBM AIF360 and diffprivlib
- References: **kamiran2012data**; **feldman2015certifying**; **calmon2017optimized**; **agrawal2000privacy**; **ji2014differential**

Constrained, Fair, Interpretable, Private or Simple Models



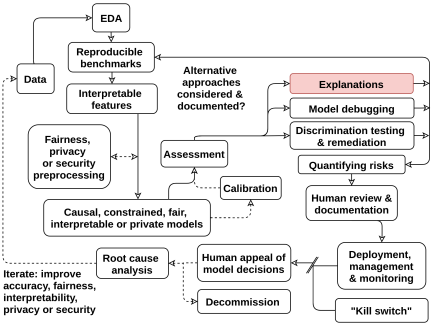
- OSS: **ga2m** (GA2M/EBM); Rudin Group models e.g. **sbrl** (SBRL); Monotonic gradient boosting machines in H2O-3 or XGBoost; pymc3
- References: **pate**; **zhang2018mitigating**; **pearl2011bayesian**; **wf_xnn** (XNN)

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

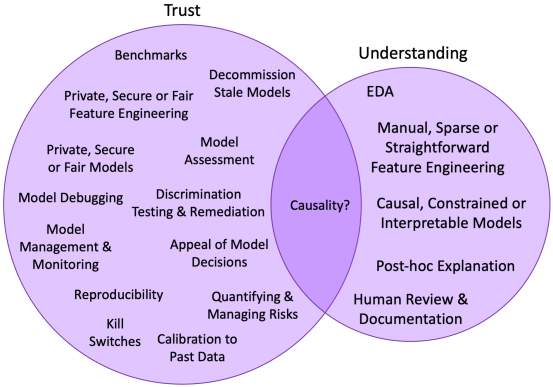


- Explanations enable *understanding* and *appeal* ... *not trust*.
- OSS: alibi, shap
- References: **wachter2017counterfactual**; **shapley**; **dt_surrogate2**; **please_stop** (criticism)

Interlude: The Time–Tested Shapley Value

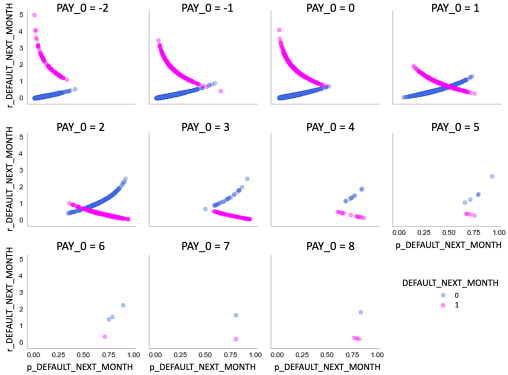
1. **In the beginning**: shapley1953value, shapley1953value
2. **Nobel-worthy contributions**: shapley1988shapley, shapley1988shapley
3. **Shapley regression**: lipovetsky2001analysis, lipovetsky2001analysis
4. **First reference in ML?** keinan2004fair, keinan2004fair
5. **Into the ML research mainstream, i.e. JMLR**: kononenko2010efficient, kononenko2010efficient
6. **Into the real-world data mining workflow ... *finally***: tree_shap, tree_shap
7. **Unification**: shapley, shapley

Interlude: Trust and Understanding

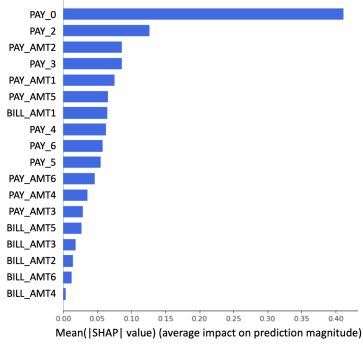


Trust and understanding in machine learning are different but complementary goals, and they are technically feasible *today*.

Interlude: Explaining Why to Distrust a Model



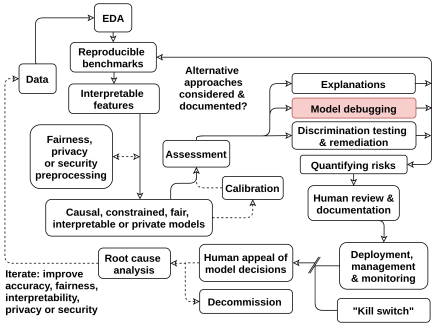
These residuals show a problematic pattern in predictions related to the most important feature, `PAY_0`.



This model over-emphasizes the most important feature, `PAY_0`.

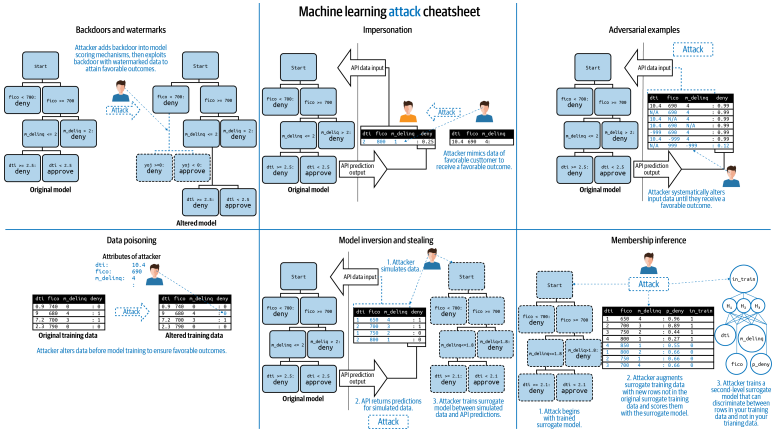
While this model is *explainable*, it's probably not *trustworthy*.

Model Debugging for Accuracy, Privacy or Security



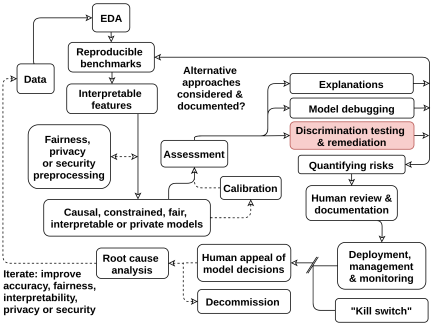
- Eliminating errors in model predictions by testing: adversarial examples, explanation of residuals, random attacks and "what-if" analysis.
- OSS: cleverhans, pdpbox, what-if tool, robustness
- References: **amershi2015modeltracker**; **papernot2018marauder**; **security_of_ml**

Machine Learning Attacks‡



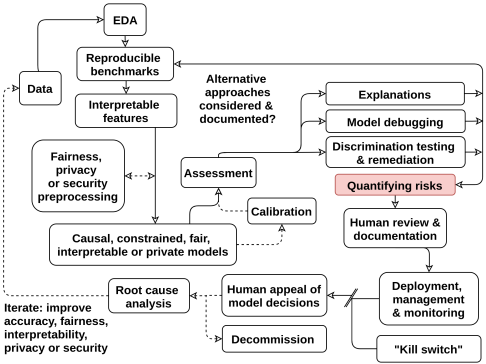
‡See https://github.com/jphall663/secure_ML_ideas for full size image and more information.

Post-hoc Disparate Impact Assessment and Remediation



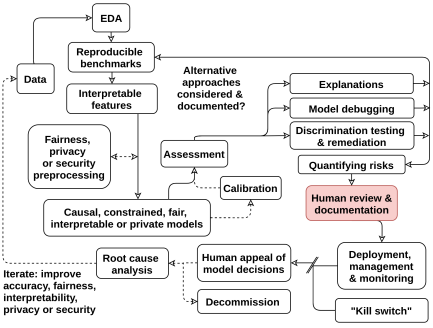
- Social bias testing should include group fairness tests and should attempt to consider individual fairness.
- OSS: aequitas, IBM AIF360, themis
- References: **dwork2012fairness**; **kamiran2012decision**; **hardt2016equality**; **feldman2015certifying**

Quantify and Plan for Risk



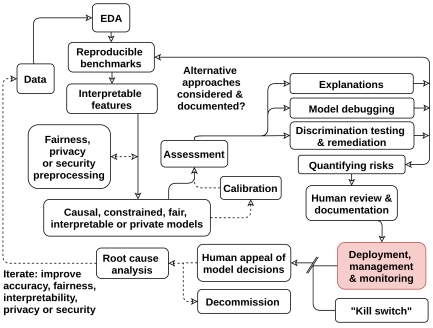
Your model will be wrong. Stake-holders need to understand and be prepared for the human and financial costs of these wrong decisions.

Human Review and Documentation



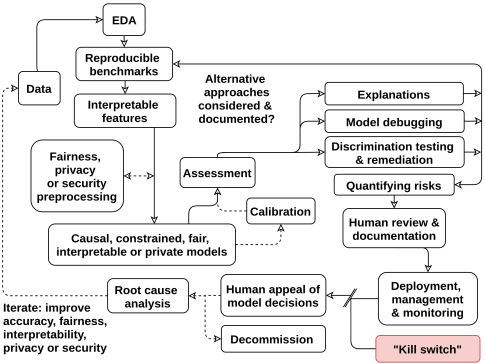
- Reference: **model_cards**
- Documentation of considered alternative approaches typically necessary for compliance.

Deployment, Management and Monitoring



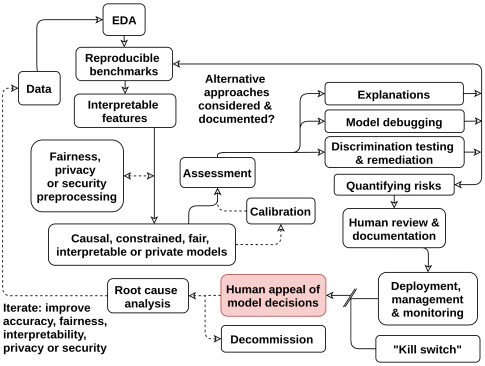
- Monitor models for accuracy, disparate impact, privacy violations or security vulnerabilities in real-time; track model and data lineage.
- OSS: DVC, gigantum, KubeFlow, mlflow, modeldb, TensorFlow ML Metadata, TensorFlow TFX, awesome-machine-learning-ops metalist
- Reference: **vartak2016m**

Kill Switches



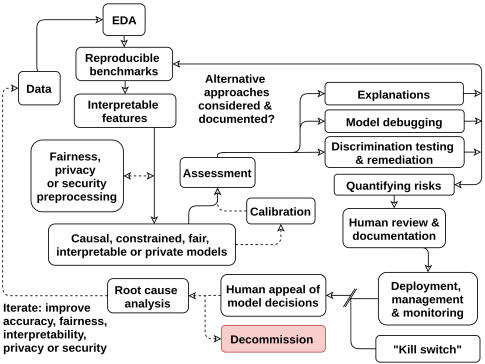
Being able to quickly turn off a misbehaving ML system is crucially important. This requires technical and organizational considerations. E.g., how much revenue is lost each minute a model is disabled?

Human Appeal



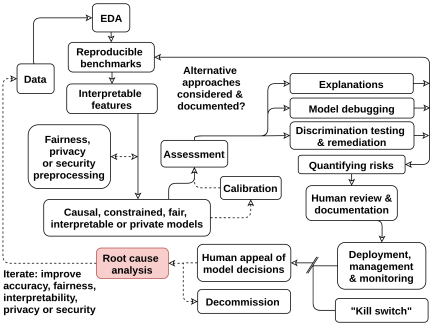
Very important, may require custom implementation for each deployment environment? Related problems exist *today*.

Decommission Model



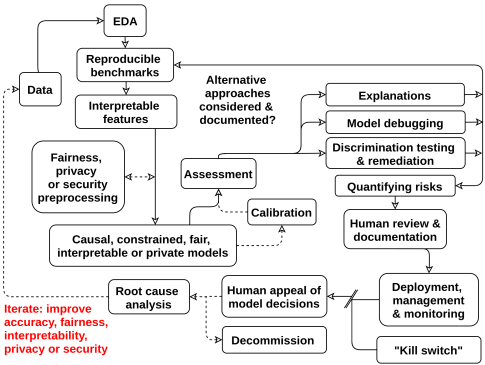
When a model becomes absolutely or relatively inaccurate, unfair, or insecure it must be taken out of service, but saved in an executable and reproducible manner.

Causality?



- Root cause analysis: can root causes be identified, verified? Formalized into model architecture?
- OSS: dowhy, pymc3
- References: **pearl2018book**; **salvatier2016probabilistic**

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.

References