# Responsible Automation: Towards Interpretable & Fair AutoML

H2O.ai

useR! 2020

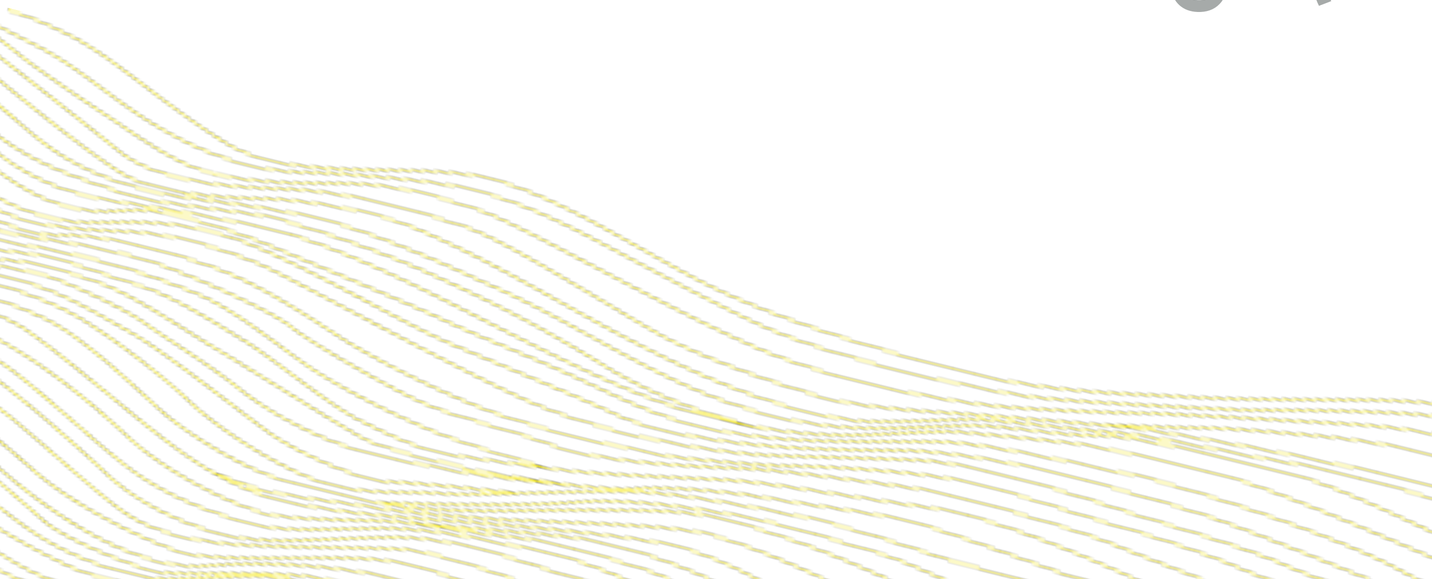
Erin LeDell Ph.D.
@ledell

# Overview

- Automatic Machine Learning (AutoML)
- Interpretable & Fair AutoML
- How & when to say “No” to ML
- What’s next?

http://github.com/ledell/useR2020-automl

# Automatic Machine Learning (AutoML)



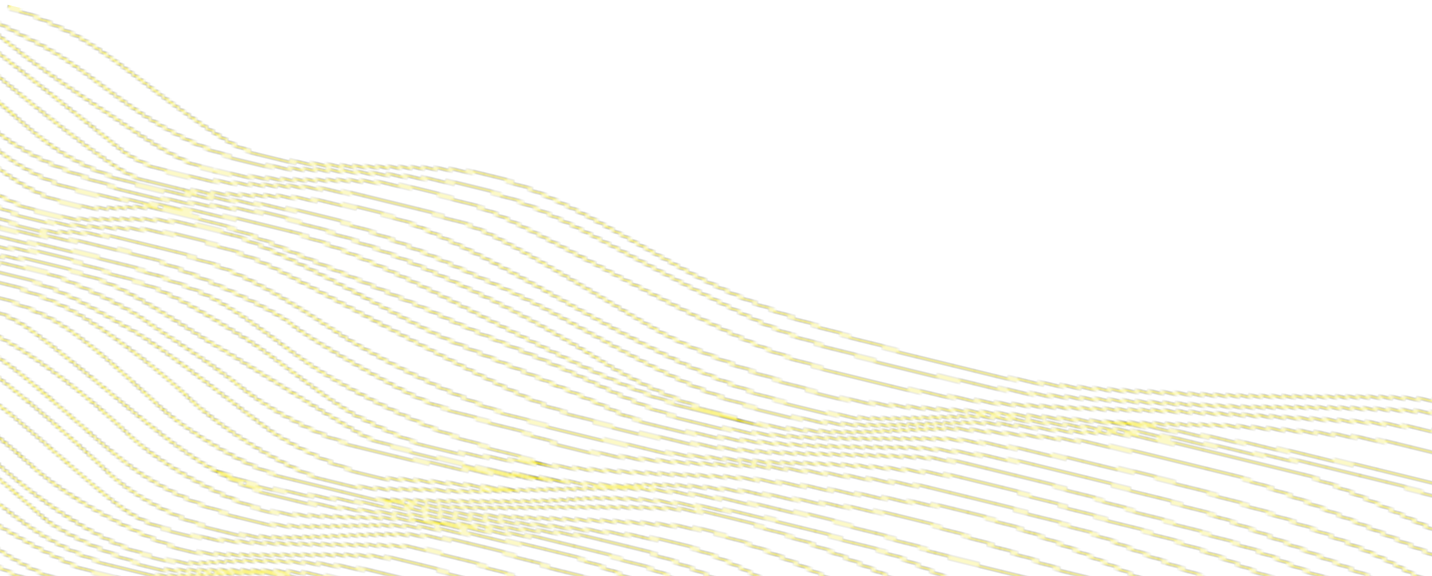
# Goals & Features of AutoML

- 🏆 Train the best model in the least amount of time.
- 📉 Reduce the human effort & expertise required in machine learning.
- 📈 Improve the performance of machine learning models.
- 🔄 Increase reproducibility & establish a baseline for scientific research or applications.

# Negative Side Effects of AutoML

- 📉 Novice data scientists can easily train models, which could increase the potential for accidental mis-use. Current AutoML systems have minimal guard rails.
- 📈 Many more models trained & deployed in prod, which means more “bad” models as well.
- 💼 AutoML can automate creating models but debugging/auditing these models is still largely a manual process. (surplus of unaudited models)

# H2O AutoML



# H2O AutoML


- Basic data pre-processing (as in all H2O algos).
- Trains a random grid of GBMs, DNNs, GLMs, etc. using a carefully chosen hyper-parameter space.
- Individual models are tuned using cross-validation.
- Two Stacked Ensembles are trained ("All Models" ensemble & a lightweight "Best of Family" ensemble).
- Returns a sorted "Leaderboard" of all models.
- All models can be easily exported to production.

# H2O AutoML in R

## Example

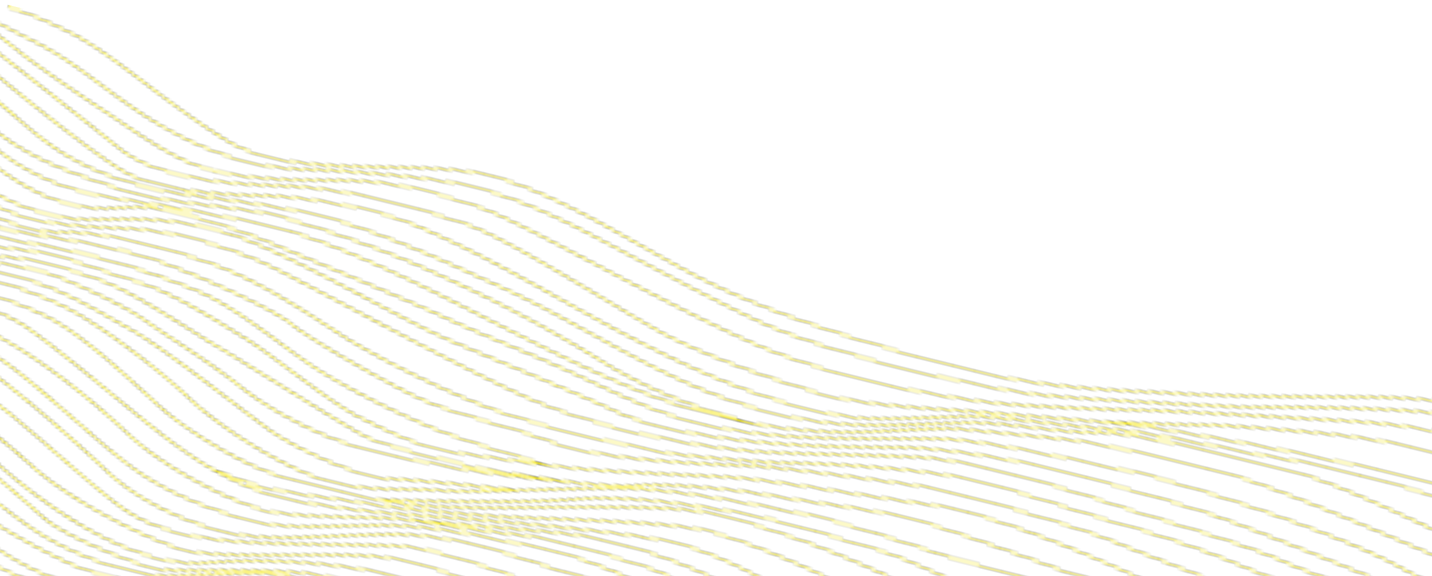
```
library(h2o)

h2o.init()

train <- h2o.importFile("train.csv")


aml <- h2o.automl(y = "response_colname",
                  training_frame = train,
                  max_runtime_secs = 600)

lb <- aml@leaderboard
```

# H2O AutoML Leaderboard

| | model_id | auc | logloss | aucpr | mean_per_class_error | rmse | mse |
|---|---|---|---|---|---|---|---|
| 1 | StackedEnsemble_AllModels_AutoML_20200709_004… | 0.8378355 | 0.2866370 | 0.4481733 | 0.2498460 | 0.2913039 | 0.08485799 |
| 2 | StackedEnsemble_BestOfFamily_AutoML_20200709_0… | 0.8369381 | 0.2869531 | 0.4462222 | 0.2500683 | 0.2914670 | 0.08495302 |
| 3 | XGBoost_3_AutoML_20200709_004415 | 0.8366588 | 0.2809896 | 0.4502926 | 0.2552901 | 0.2894478 | 0.08378002 |
| 4 | GBM_4_AutoML_20200709_004415 | 0.8330289 | 0.2848382 | 0.4239271 | 0.2593298 | 0.2919957 | 0.08526147 |
| 5 | GBM_3_AutoML_20200709_004415 | 0.8325824 | 0.2852444 | 0.4195761 | 0.2552272 | 0.2922670 | 0.08542002 |
| 6 | GBM_2_AutoML_20200709_004415 | 0.8323248 | 0.2855498 | 0.4185351 | 0.2589230 | 0.2924915 | 0.08555129 |
| 7 | GBM_1_AutoML_20200709_004415 | 0.8322315 | 0.2855884 | 0.4200573 | 0.2622791 | 0.2922375 | 0.08540278 |
| 8 | XGBoost_1_AutoML_20200709_004415 | 0.8317490 | 0.2858897 | 0.4326282 | 0.2618297 | 0.2923182 | 0.08544993 |
| 9 | GBM_5_AutoML_20200709_004415 | 0.8296069 | 0.2874258 | 0.4040567 | 0.2569593 | 0.2938664 | 0.08635746 |
| 10 | XGBoost_2_AutoML_20200709_004415 | 0.8277037 | 0.2899311 | 0.4265391 | 0.2624847 | 0.2943874 | 0.08666391 |
| 11 | DRF_1_AutoML_20200709_004415 | 0.8120043 | 0.3008964 | 0.3722857 | 0.2731671 | 0.2991530 | 0.08949252 |
| 12 | GLM_1_AutoML_20200709_004415 | 0.6873574 | 0.3510707 | 0.2172795 | 0.3673990 | 0.3194751 | 0.10206432 |

Example Leaderboard for binary classification

# Interpretable & Fair AutoML



# Interpretable ML & Algorithmic Fairness in R

🔬 Interpretability in R: `iml`, `DALEX`, `lime`, etc.

🌊 H2O integrations & demos:

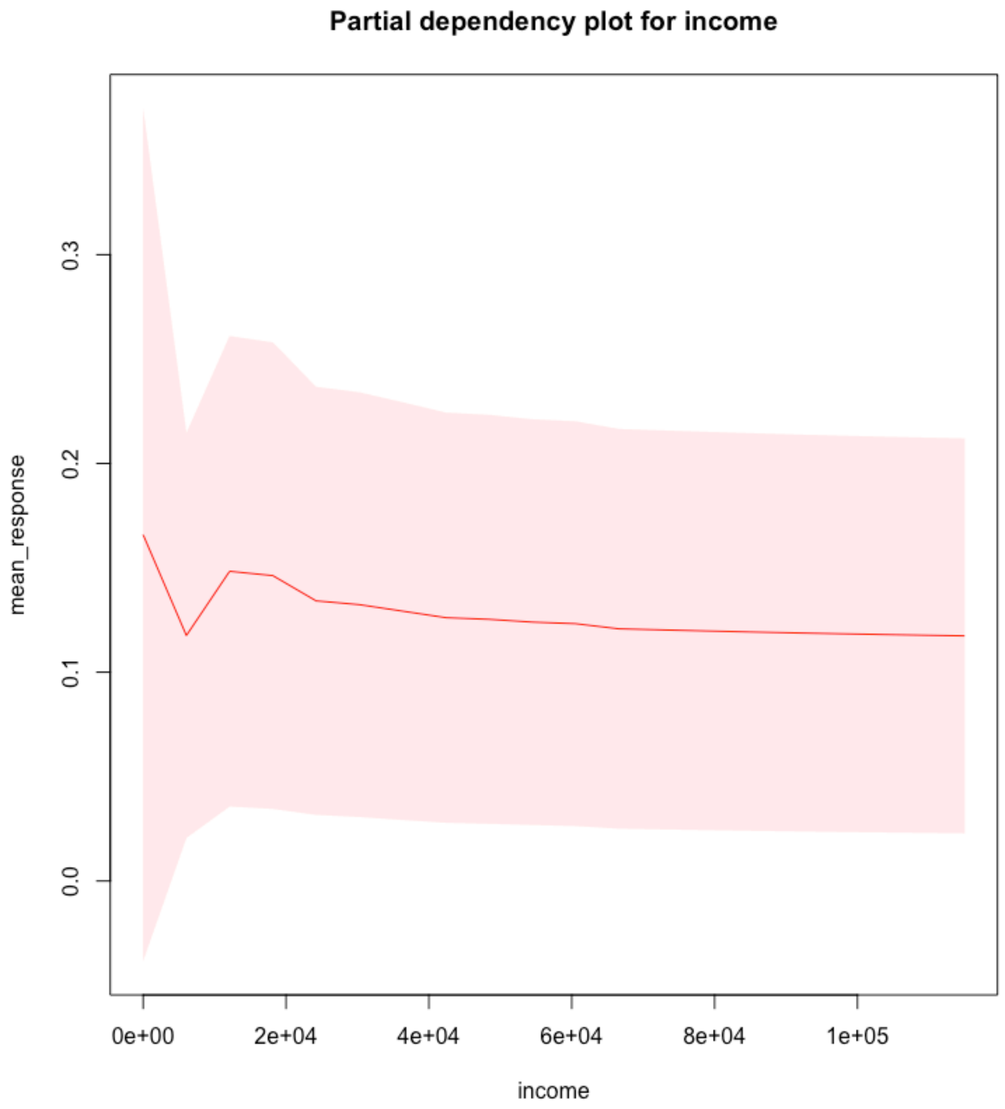
`iml`: http://uc-r.github.io/iml-pkg

`DALEX`: http://uc-r.github.io/dalex

`lime`: https://github.com/woobe/useR2019_h2o_tutorial

⚖️ Fairness in R: `fairness` (only one?)

# Partial Dependence Plots (PDP)




The partial dependence plot (PDP) shows the marginal effect one or two features have on the predicted outcome of a machine learning model.

(J.H. Friedman, 2001)

# Fairness in ML

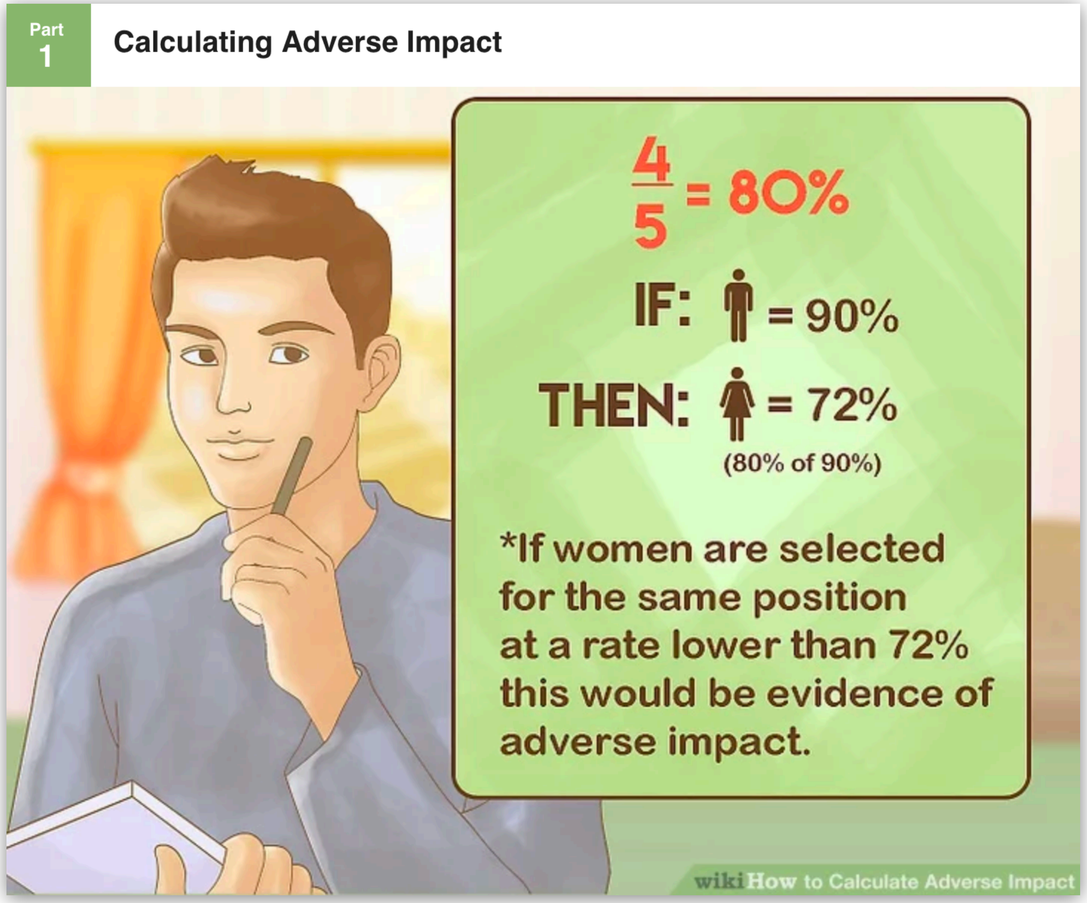
What is “fairness” in ML? ⚖️

- Very hard question to answer. 🤔
- Describes a broad set of problems, not a specific approach or metric.

Why it’s hard:

- Individual fairness vs. group fairness
- Law/policy is somewhat fuzzy – moving target…

# Disparate Impact Analysis




## Popular disparity metrics

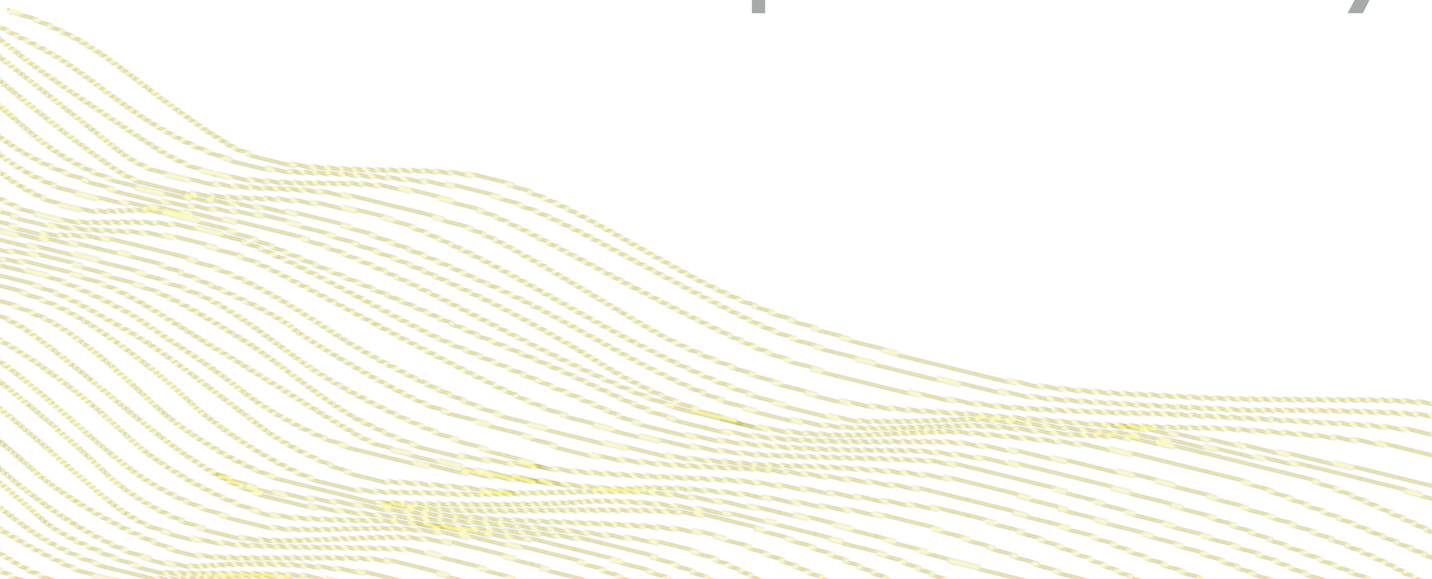
- Classification: Adverse Impact Ratio (AIR), Marginal Effect (ME)
- Regression: Standardized Mean Difference (SMD)

https://www.wikihow.com/Calculate-Adverse-Impact

# Remediation

- Use Disparate Impact Analysis to choose a model that's fair (based on metrics like AIR, SMD, etc.)
- If there are no fair models, then: 🔧
  - Fix data: row weights or up/down sampling
  - Fix model: regularize impact of demographic features; adversarial debiasing.
  - Fix predictions: re-weight preds to achieve fairness
  - Give up!

# H2O AutoML
# Interpretability & Fairness



# H2O AutoML Interpretability

Work-in-progress: 🔬 🛠️

- Feature importance comparisons
- TreeSHAP for tree-based models (global & local)
- Partial dependence (PD) plots
- Individual Conditional Expectation (ICE) plots
- Constrained (monotonic) models
- Disparate Impact Analysis
- Fairness metrics (AIR, ME, SMD)

# H2O AutoML Fairness Demo

🏘️ Home Mortgage Disclosure Act (HMDA) data

❓ Predict “high-priced” loan

```
# Calculate any adverse impact across subgroups for all models
da <- h2o.get_disparate_analysis(aml = aml,
                                 newdata = test,
                                 columns = c("derived_sex", "derived_race"),
                                 favorable_class = "0")
```

```
# Calculate adverse impact across subgroups for leader model
dm <- calculate_disparate_measures(model = aml@leader,
                                   newdata = test,
                                   columns = c("derived_sex", "derived_race"),
                                   favorable_class = "0")
```

http://github.com/ledell/useR2020-automl

# H2O AutoML Leaderboard + Fairness

| | model_id | auc | air_min | air_mean | air_median | adverse_impact |
|---|---|---|---|---|---|---|
| 1 | StackedEnsemble_AllModels_AutoML_20200709_224... | 0.8373151 | 0.6287504 | 0.8645903 | 0.8685584 | TRUE |
| 2 | StackedEnsemble_BestOfFamily_AutoML_20200709_2... | 0.8367662 | 0.6211536 | 0.8626577 | 0.8696150 | TRUE |
| 3 | XGBoost_3_AutoML_20200709_224644 | 0.8362910 | 0.6561777 | 0.8758646 | 0.8854495 | TRUE |
| 4 | GBM_3_AutoML_20200709_224644 | 0.8334076 | 0.6750566 | 0.8887162 | 0.9061072 | TRUE |
| 5 | GBM_4_AutoML_20200709_224644 | 0.8330067 | 0.6945701 | 0.8884441 | 0.8999003 | TRUE |
| 6 | GBM_2_AutoML_20200709_224644 | 0.8318526 | 0.6375823 | 0.8664354 | 0.8674385 | TRUE |
| 7 | XGBoost_1_AutoML_20200709_224644 | 0.8314702 | 0.6944284 | 0.8932051 | 0.9031095 | TRUE |
| 8 | GBM_1_AutoML_20200709_224644 | 0.8308022 | 0.6723884 | 0.8825634 | 0.8872850 | TRUE |
| 9 | GBM_5_AutoML_20200709_224644 | 0.8291239 | 0.6656952 | 0.8860885 | 0.9048964 | TRUE |
| 10 | XGBoost_2_AutoML_20200709_224644 | 0.8269975 | 0.7096087 | 0.9003328 | 0.9055053 | TRUE |
| 11 | DRF_1_AutoML_20200709_224644 | 0.8149284 | 0.6212269 | 0.8458582 | 0.8720485 | TRUE |
| 12 | GLM_1_AutoML_20200709_224644 | 0.6806845 | 0.5437908 | 0.8022767 | 0.7945545 | TRUE |

Example Leaderboard (fairness metrics added, other columns dropped)

# Disparate Measures Info (AutoML leader)

| | derived_sex | derived_race | reference | Total | Selected | air | adverse_impact |
|---|---|---|---|---|---|---|---|
| 1 | Female | *NA* | 2 | 7151 | 5698 | 0.9729337 | FALSE |
| 2 | Male | *NA* | 2 | 11413 | 9347 | 1.0000000 | FALSE |
| 3 | *NA* | 2 or more minority races | 9 | 32 | 26 | 0.8614781 | FALSE |
| 4 | Female | 2 or more minority races | 11 | 16 | 13 | 0.8569259 | FALSE |
| 5 | Male | 2 or more minority races | 11 | 16 | 13 | 0.8569259 | FALSE |
| 6 | *NA* | American Indian or Alaska Native | 9 | 119 | 87 | 0.7751632 | FALSE |
| 7 | Female | American Indian or Alaska Native | 11 | 51 | 31 | 0.6410788 | FALSE |
| 8 | Male | American Indian or Alaska Native | 11 | 68 | 56 | 0.8685584 | FALSE |
| 9 | *NA* | Asian | 9 | 1284 | 1211 | 1.0000000 | FALSE |
| 10 | Female | Asian | 11 | 416 | 388 | 0.9836901 | FALSE |
| 11 | Male | Asian | 11 | 868 | 823 | 1.0000000 | FALSE |
| 12 | *NA* | Black or African American | 9 | 1773 | 1136 | 0.6793451 | TRUE |
| 13 | Female | Black or African American | 11 | 884 | 527 | 0.6287504 | TRUE |
| 14 | Male | Black or African American | 11 | 889 | 609 | 0.7224960 | TRUE |
| 15 | *NA* | Native Hawaiian or Other Pacific Islander | 9 | 58 | 51 | 0.9323158 | FALSE |
| 16 | Female | Native Hawaiian or Other Pacific Islander | 11 | 22 | 20 | 0.9587982 | FALSE |
| 17 | Male | Native Hawaiian or Other Pacific Islander | 11 | 36 | 31 | 0.9081950 | FALSE |
| 18 | *NA* | Race Not Available | 9 | 1183 | 988 | 0.8855092 | FALSE |
| 19 | Female | Race Not Available | 11 | 472 | 393 | 0.8781535 | FALSE |
| 20 | Male | Race Not Available | 11 | 711 | 595 | 0.8826068 | FALSE |
| 21 | *NA* | White | 9 | 14115 | 11546 | 0.8673044 | FALSE |
| 22 | Female | White | 11 | 5290 | 4326 | 0.8624834 | FALSE |
| 23 | Male | White | 11 | 8825 | 7220 | 0.8628640 | FALSE |

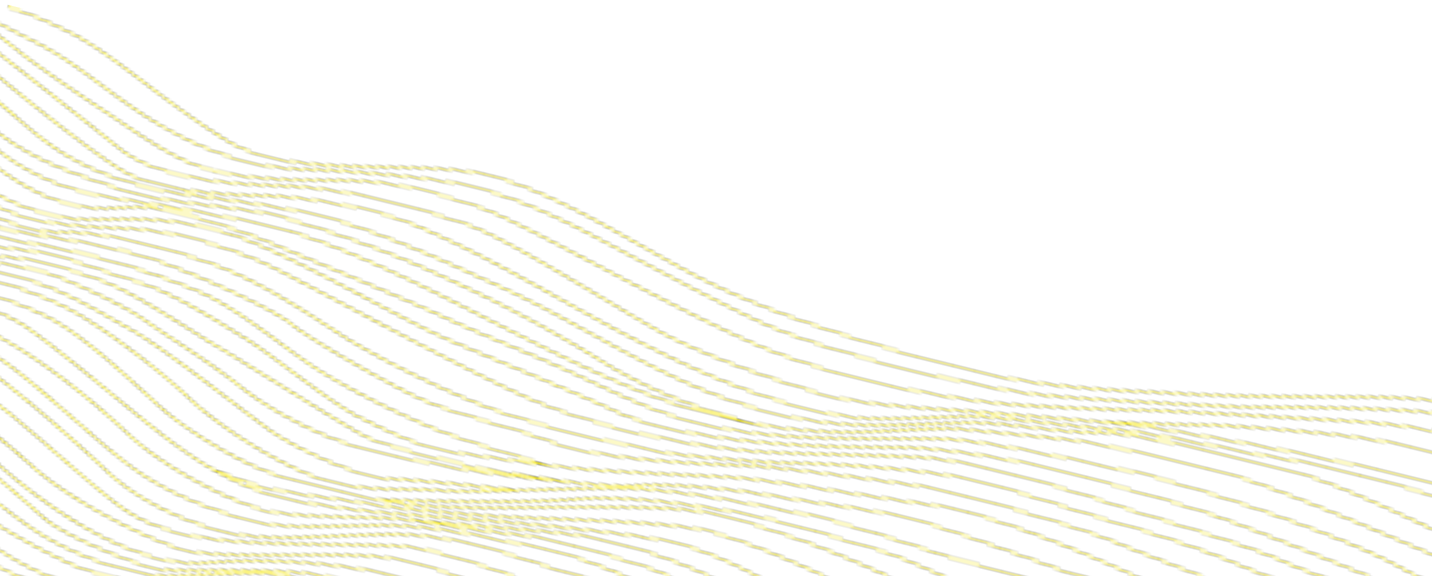
Adverse Impact detected!

This model discriminates against black people.

❌🚮

Example Disparate Measures Data computed on a single model (columns subsetted)

# How and When to "ML"



# Don’t miss the forest for the trees...

"Can I minimize differences in accuracy between subgroups" is less important than "should this be built at all"

@rctatman

Dr. Rachael Tatman

http://www.rctatman.com/talks/what-i-wont-build

# When to use ML?

Considerations can be ethical and/or technical.
Does your ML model:

- 🚫 Harm people (esp. marginalized groups)?
- 🚫 Shift power (esp. to the majority)?
- 🚫 Fail on outliers?
- 🚫 Have strong confidence about wrong answers?

# When to use ML?

Timnit Gebru

A lot of times, people are talking about bias in the sense of equalizing performance across groups. They're not thinking about the underlying foundation, whether a task should exist in the first place, who creates it, who will deploy it on which population, who owns the data, and how is it used?

The root of these problems is not only technological. It's social. Using technology with this underlying social foundation often advances the worst possible things that are happening. In order for technology not to do that, you have to work on the underlying foundation as well. You can't just close your eyes and say: "Oh, whatever, the foundation, I'm a scientist. All I'm going to do is math."

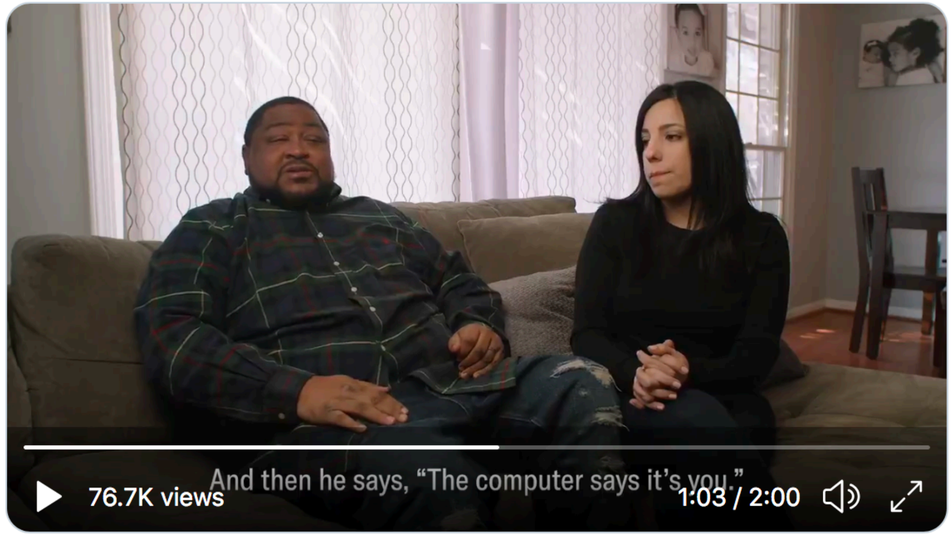
https://www.nytimes.com/2019/11/19/technology/artificial-intelligence-bias.html

# Unethical ML: Law Enforcement

ACLU
@ACLU

BREAKING: We're filing a complaint against Detroit police for wrongfully arresting Robert Williams, an innocent Black man — all because face recognition technology can't tell Black people apart.

Officers hauled him away in front of his kids and locked him up for 30 hours.




4:00 AM · Jun 24, 2020 · Twitter Media Studio

**1.8K** Retweets and comments **3.9K** Likes

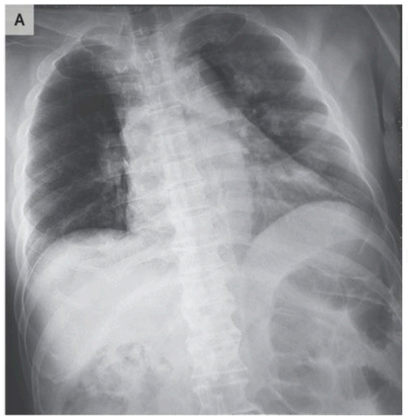
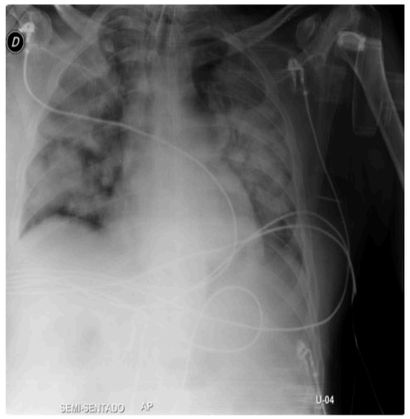
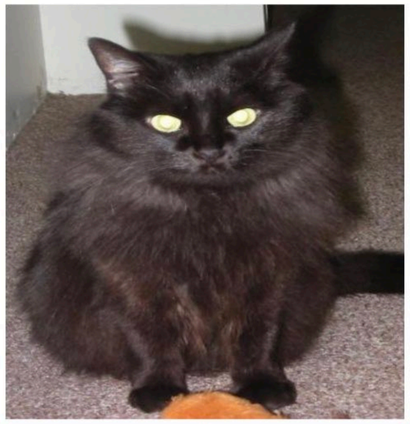
ajl.org

“The computer says it’s you.”

😱 🚫

https://twitter.com/ACLU/status/1275745489544081408

# Bad ML: #covidcat

| True Label | COVID-19 (Training Data) | COVID-19 (Unseen Data) | Cat (Unrelated Data) |
|---|---|---|---|
| |  |  |  |

| Model | Prediction | Confidence | Prediction | Confidence | Prediction | Confidence |
|---|---|---|---|---|---|---|
| DNN | COVID-19 | 99.7% | Non-COVID | 75.1% | COVID-19 | 100% |
| BNN | COVID-19 | 95.5% | COVID-19 | 67.1% | COVID-19 | 99.8% |
| Ours | COVID-19 | 99.9% | COVID-19 | 69.0% | COVID-19 | 50.1% |

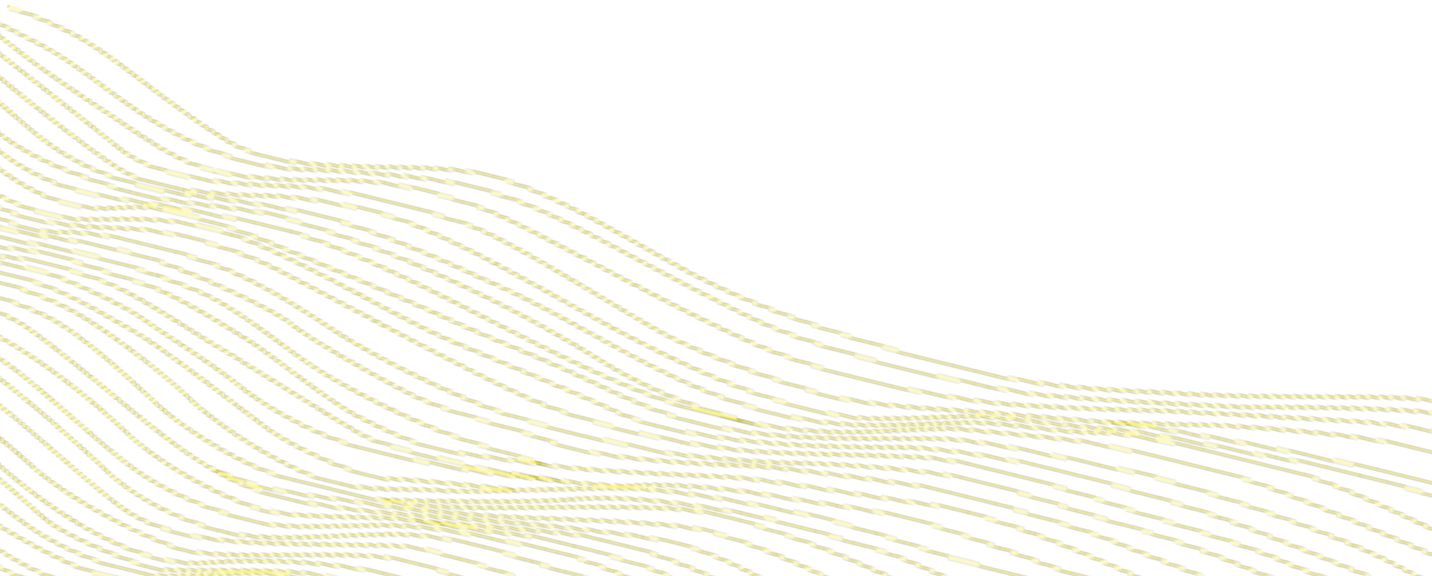
*Figure 1.* The predictions of popular deep learning models that are trained to detect the presence of COVID-19 in chest X-Ray images


**Abstract**

Deep Neural Networks (DNNs) are known to make highly overconfident predictions on Out-of-Distribution data. Recent research has shown that uncertainty-aware models, such as, Bayesian Neural Network (BNNs) and Deep Ensembles, are less susceptible to this issue. However research in this area has been largely confined to the big data setting. In this work, we show that even state-of-the-art BNNs and Ensemble models tend to make overconfident predictions when the amount of training data is insufficient. This is especially concerning for emerging applications in physical sciences and healthcare where overconfident and inaccurate predictions can lead to disastrous consequences. To address the issue of accurate uncertainty (or confidence) estimation in the small-data regime, we propose a probabilistic generalization of the popular sample-efficient non-parametric kNN approach. We demonstrate the usefulness of the proposed approach on a real-world application of COVID-19 diagnosis from chest X-Rays by (a) highlighting surprising failures of existing techniques, and (b) achieving superior uncertainty quantification as compared to state-of-the-art.


https://tinyurl.com/icml-covidcat

# What’s next for ML?



# What's next for ML?

Traditional ML: A "prediction-only" technology that has served the tech and eCommerce industry very well.

Next-generation ML: Needs to support critical scientific applications (healthcare, etc.) which require prediction, as well as understanding of the underlying stochastic phenomena.

# What's next for ML?

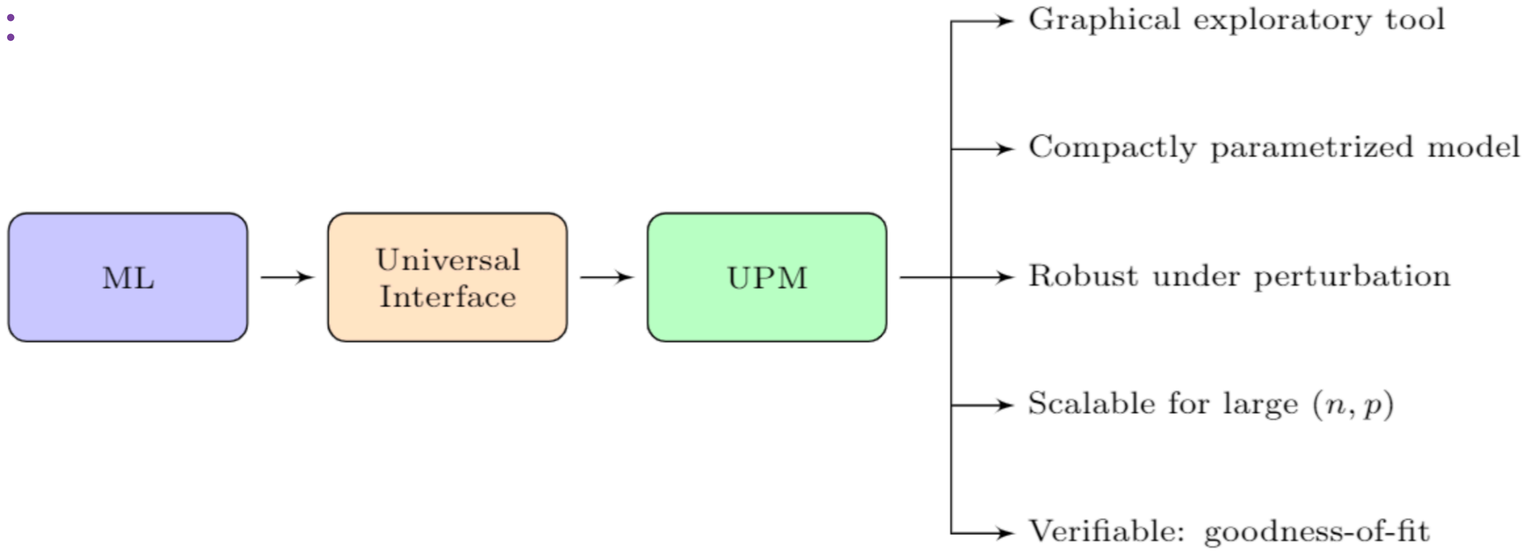
Five main building blocks of ML 2.0:

- Automation – AutoML.
- Exploration – Can we see if ML captured the important information?
- Prediction – Uncertainty distribution prediction machine (beyond point estimation).
- Attribution – Identifying features that are most predictive for the response (beyond mean predictors).
- Verification – Check if model is congruent with data.

# Universal Learning Architecture

## Universal Prediction Machine (UPM):

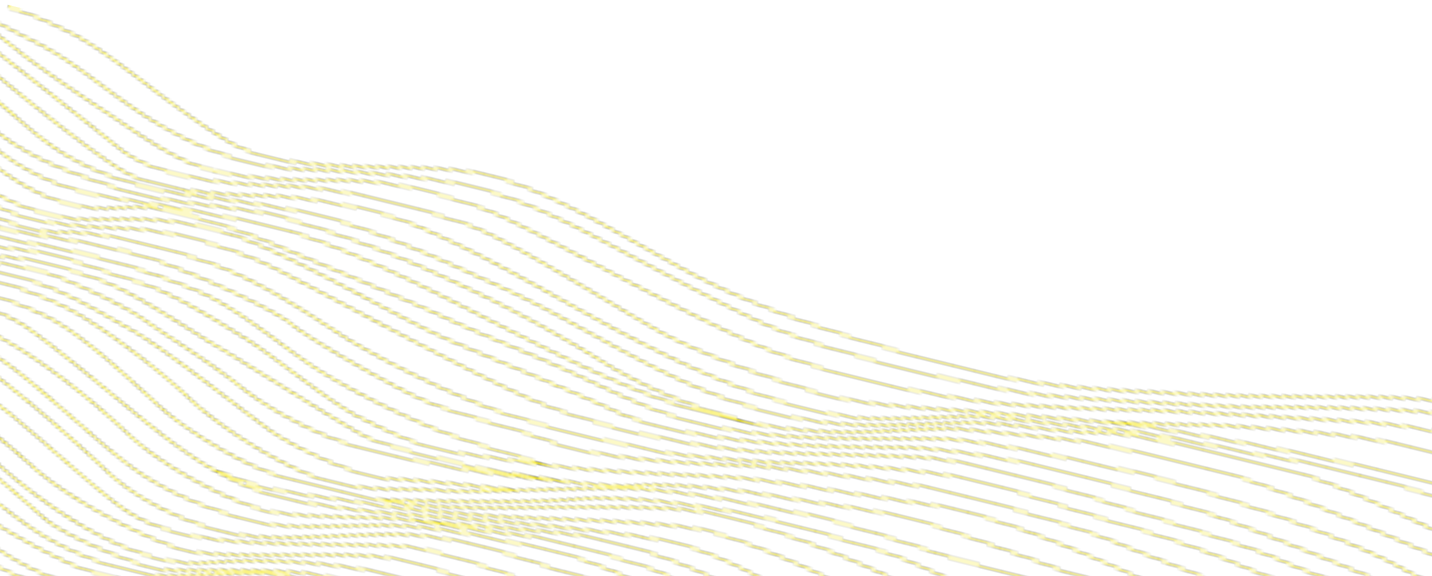
An interface which can convert any ML algorithm into an "Exploratory Learning Machine".




Collaboration with Deep Mukhopadhyay (Stanford)


- *Mukhopadhyay, S., and Wang, K (2020) Breiman's "Two Cultures" Revisited and Reconciled. arXiv:2005.13596*
- *Mukhopadhyay, S., LeDell, E., and Wilkinson, L. (2020) Next-Generation Integrated Learning Machine, H2O.ai report (unreleased)*

# Resources



# Learn H2O AutoML!

- Docs: https://tinyurl.com/h2o-automl-docs
- AutoML R tutorials: https://tinyurl.com/h2o-automl-tutorials
- LatinR H2O tutorial: https://tinyurl.com/latinr-h2o

# Thank you!

@ledell on Github, Twitter
erin@h2o.ai