# Explainable AI with H2O Driverless AI's machine learning interpretability module

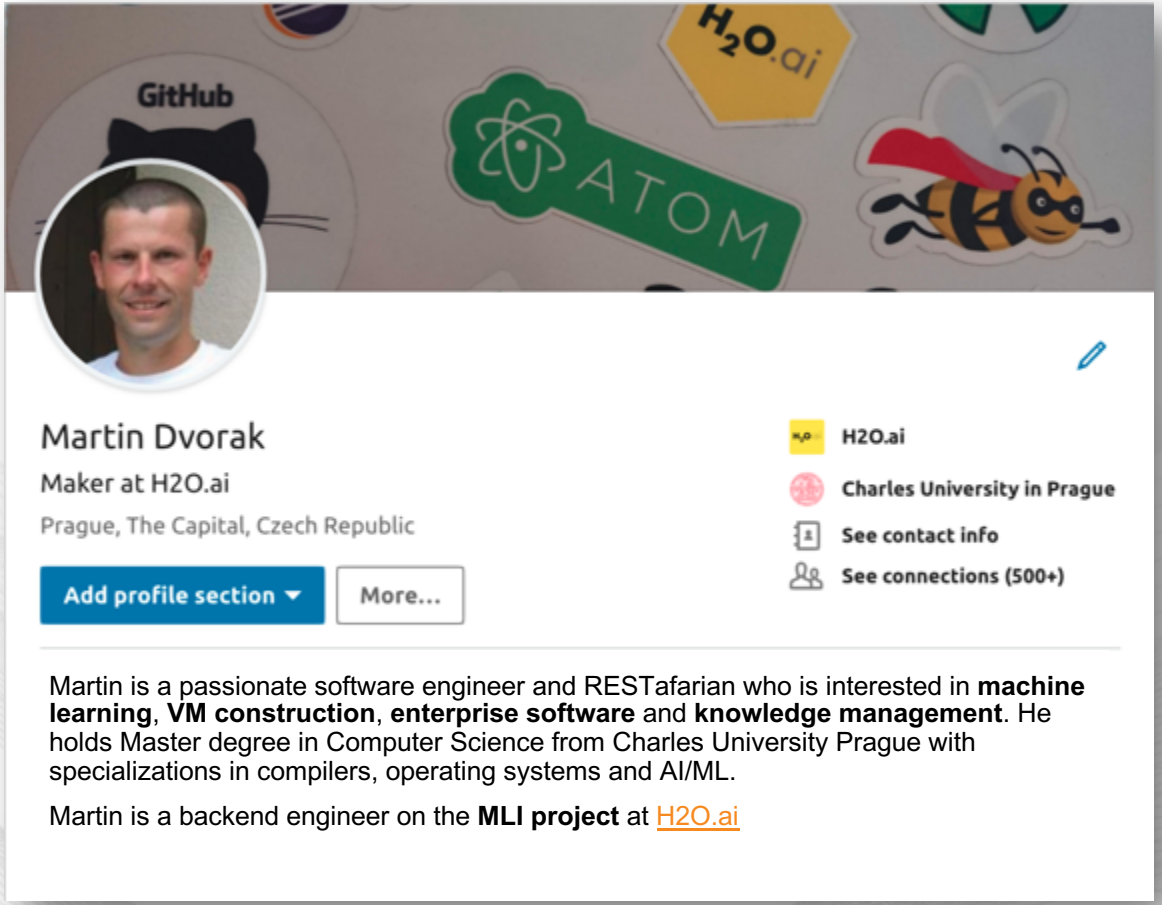
- Explainable AI is in the news, and for good reason. Financial services companies have cited the ability to explain AI-based decisions as one of the critical roadblocks to further adoption of AI for their industry. Transparency, accountability, and trustworthiness of data-driven decision support systems based on AI and machine learning are serious regulatory mandates in banking, insurance, healthcare, and other industries. From pertinent regulations, to increasing customer trust, data scientists and business decision makers must show AI-based decisions can be explained.
- H2O Driverless AI does explainable AI today with its machine learning interpretability (MLI) module. This capability in H2O Driverless AI employs a unique combination of techniques and methodologies to explain the results of both Driverless AI models and external models.

# Explainable AI with H2O Driverless AI's machine learning interpretability module

Martin Dvorak
Software Engineer, H2O.ai
martin.dvorak@h2o.ai

H2O.ai Prague Meetup #3 2019/5/16

# ABOUT ME


GitHub
H2O.ai
ATOM
Martin Dvorak
Maker at H2O.ai
Prague, The Capital, Czech Republic
Add profile section
More...
H2O.ai
Charles University in Prague
See contact info
See connections (500+)
Martin is a passionate software engineer and RESTafarian who is interested in **machine learning**, **VM construction**, **enterprise software** and **knowledge management**. He holds Master degree in Computer Science from Charles University Prague with specializations in compilers, operating systems and AI/ML.
Martin is a backend engineer on the **MLI project** at H2O.ai

# AGENDA

- **Intro**
  - Context and scope.
- **Why**
  - Explainability matters.
- **What**
  - Steps to build human-centered, low-risk models.
- **How**
  - Explaining models using of H2O.ai's solution.

# Intro

*Terminology, scope and context*

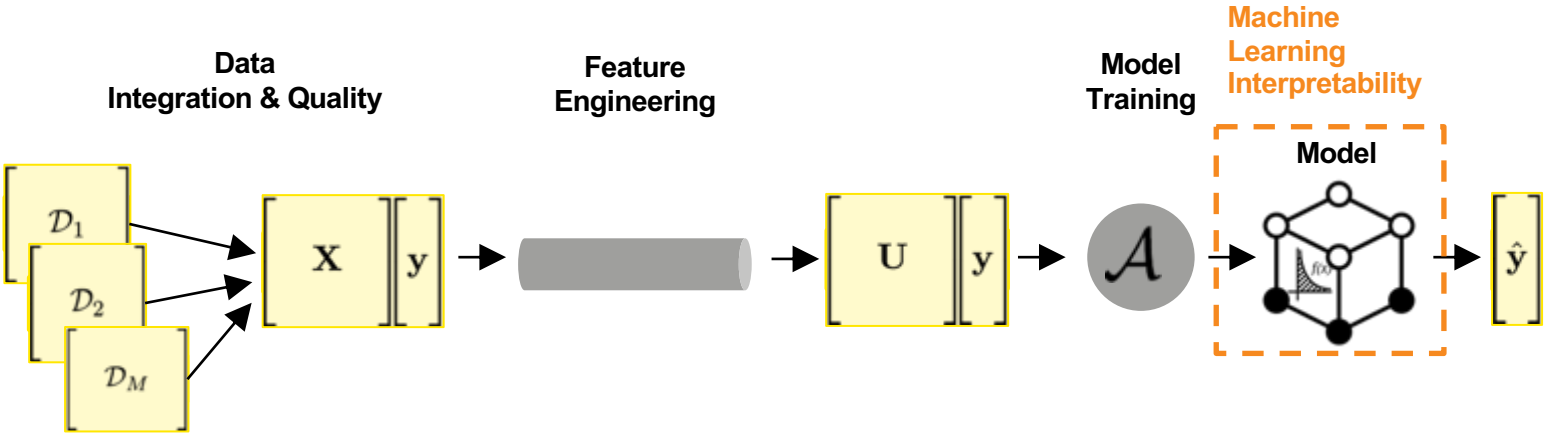
# Terminology, Scope and Context

INTRO

- **Machine Learning Interpretability**
  - *"[Machine learning interpretability] is the ability to explain or present in understandable terms to a human."* –https://arxiv.org/pdf/1702.08608.pdf
- **Structured data**
  - No ~~image~~, ~~video~~ and ~~sound~~ > deep learning typically not used.
  - Tabular data and supervised ML.
- **Auto ML**
  - H2O Driverless AI (DAI) product (not OSS).
- **MLI module**
  - Solution based on MLI module of H2O Driverless AI.

# Terminology, Scope and Context

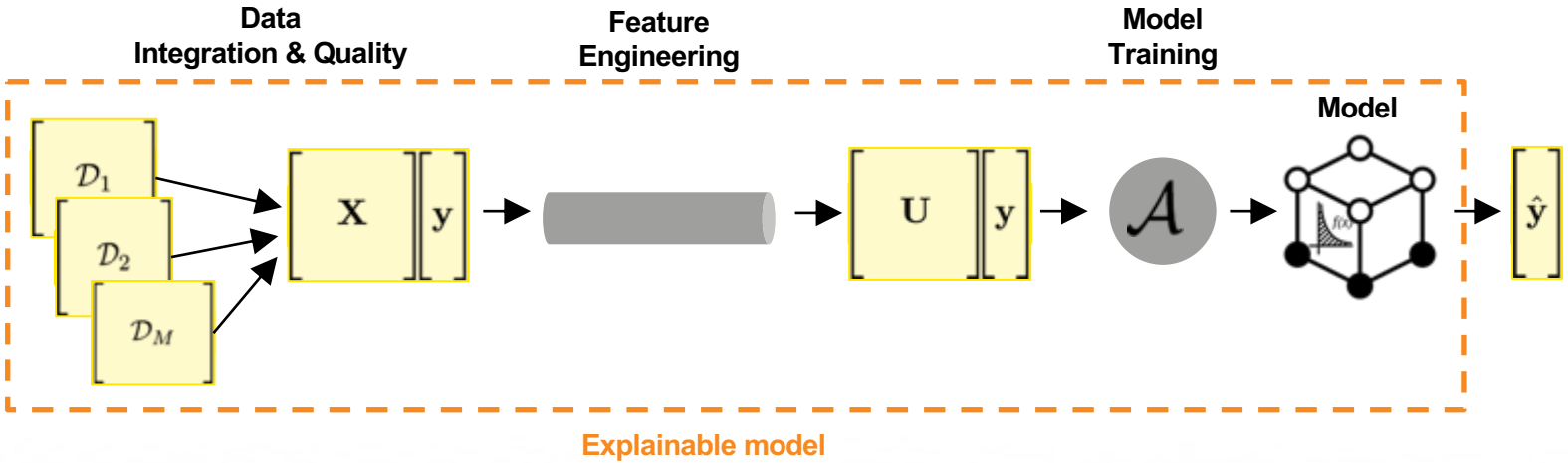
INTRO




End to end

# Terminology, Scope and Context

INTRO




End to end

# **Why** explainability matters

*Problem statement*

# Potential Performance and Interpretability **Trade-off**

(Trade-off)

White box model

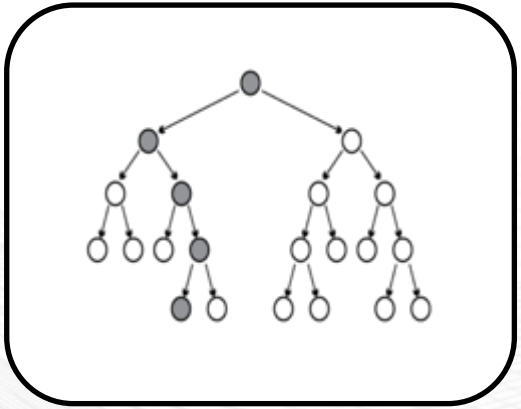
Black box model

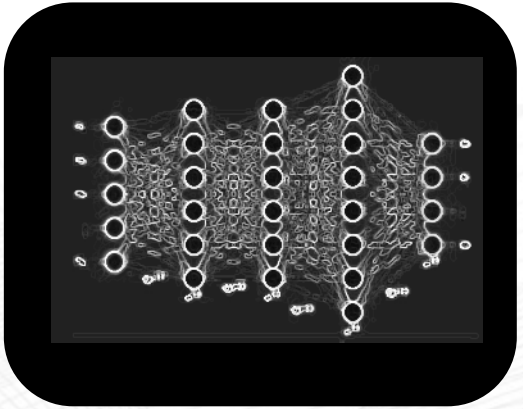
Feature engineering + Algorithm(s)

Balance

# Potential Performance and Interpretability **Trade-off**

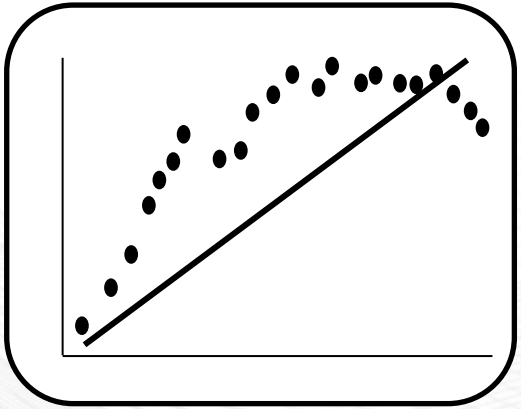
(Trade-off)

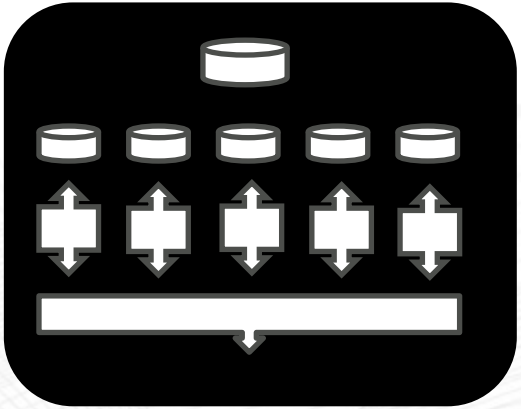




# Potential Performance and Interpretability **Trade-off**

(Trade-off)





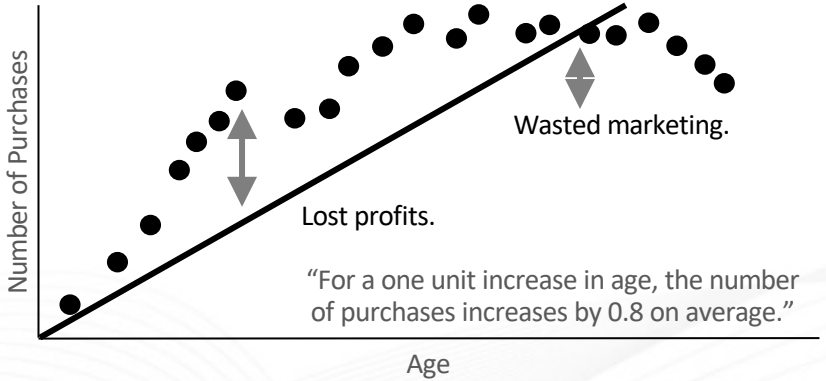
(Trade-off)

# Potential Performance and Interpretability **Trade-off**

H2O.ai

***Exact*** explanations for ***approximate*** models.

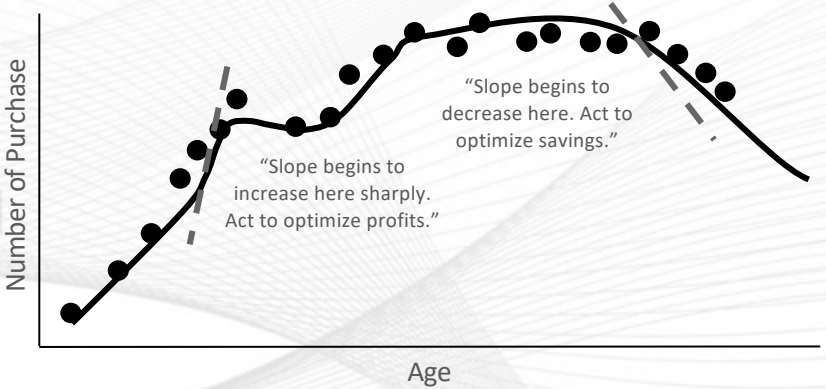
Linear models




***Approximate*** explanations for ***exact*** models.

Sometimes…

Machine learning models



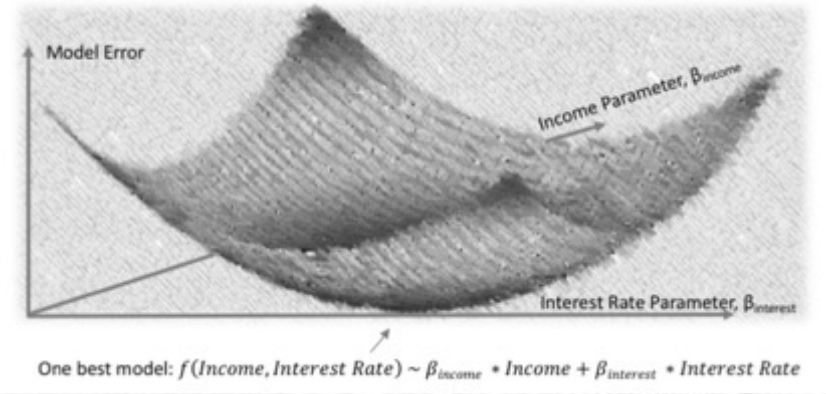
Trade-off

Multiplicity

# **Multiplicity** of Good Models

H$_2$O.ai

- For a given well-understood dataset there is usually **one** best linear model, but…

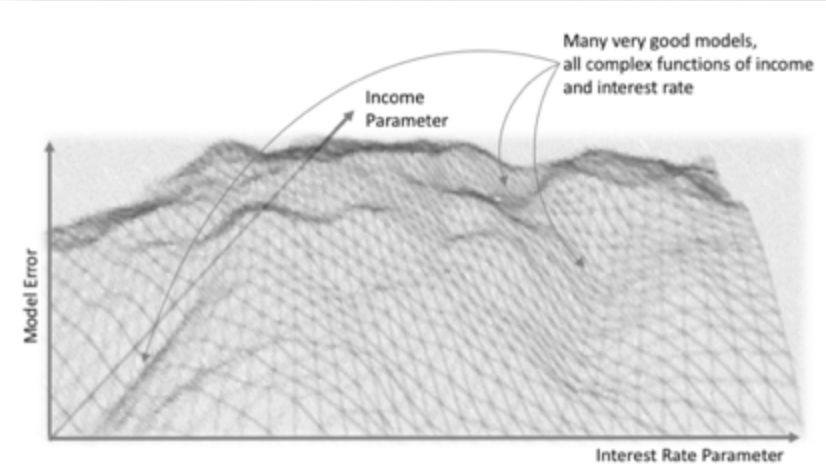


Trade-off

Multiplicity

# **Multiplicity** of Good Models

- … for a given well-understood dataset there are usually **many good** ML models. Which one to **choose**?
- Same **objective metrics** values, **performance**, …
- This is often referred to as “the **multiplicity** of good models.” -- Leo Breiman



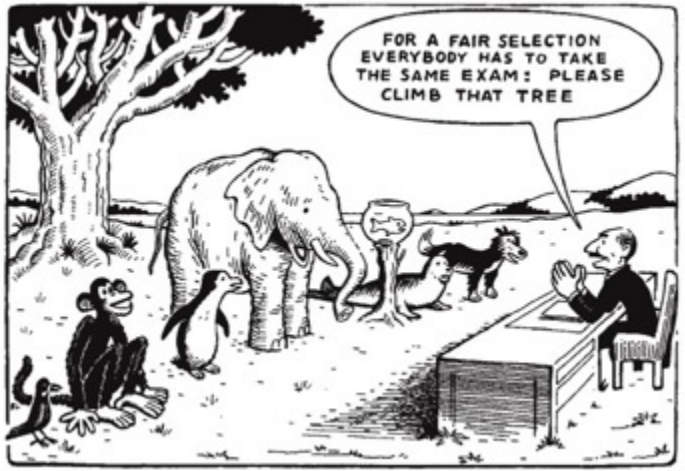
Trade-off

Multiplicity

Fairness

# **Fairness** and Social Aspects

- Gender
- Age
- Ethnicity
- Health
- Sexual behavior




- Avoid **discriminatory models** and remediate disparate impact.

Trade-off

Multiplicity

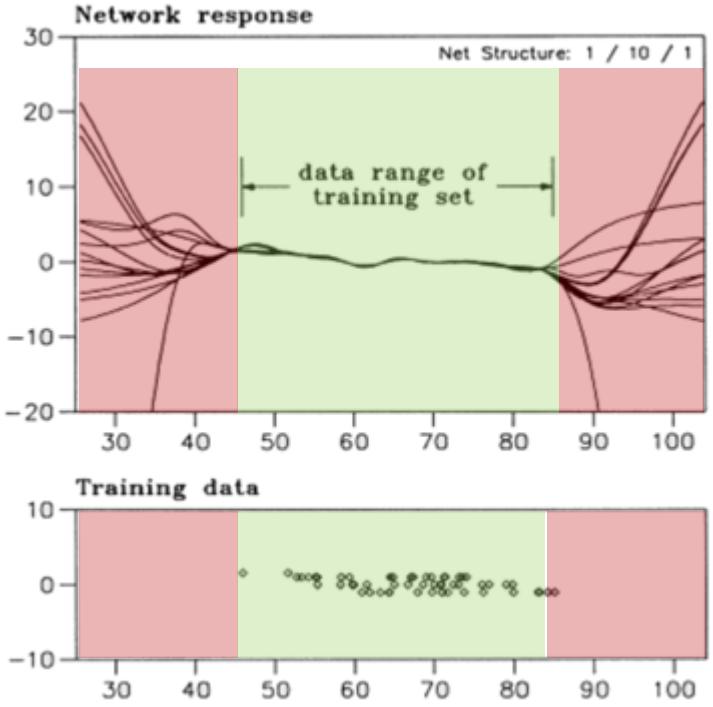
Fairness

Trust

# **Trust** of model producers & consumers

- Dataset
  vs.
  **real world**
- ML **adoption**

- Introspection
- Sensitivity
- OOR
- Diagnostics
- “Debugging”




Source: http://www.vias.org/tmdatanaleng/

Trade-off

Multiplicity

Fairness

Trust

Security

# **Security** and **Hacking**

H2O.ai

- Goal: **compromise** model integrity
- Attack types:
  - **Exploratory**
    - Surrogate model trained to identify vulnerabilities ~ MLI.
    - Trial and error (for specific class) x indiscriminate attacks.
  - **Causative**
    - Models trained w/ adversary datasets.
    - Local model > adversarial instance > target model.
      - Standard / continuous learning.
    - **Integrity** (compromise system integrity)
      - False negative instance e.g. fraud passes check.
    - **Availability** (compromise system availability)
      - False positive instance e.g. blocks access to legitimate instances.

Trade-off

Multiplicity

Fairness

Trust

Security

Regulation

# **Regulated & Controlled** Environments

H2O.ai

- Legal requirements
  - Banking, insurance, healthcare, …
- Predictions explanation
  - Decisions justification (reason codes*, …).
- Fairness
- Security
- Accuracy first vs. **interpretability** first
  - Competitions vs. real world.

Trade-off

Multiplicity

Fairness

Trust

Security

Regulation

# Explainability Matters

H2O.ai

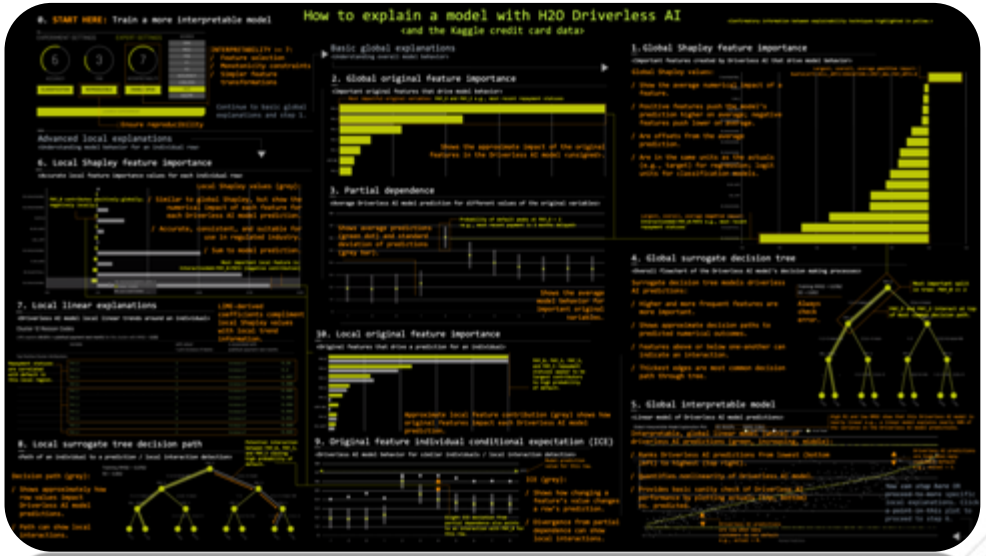
- **Balance** Performance and interpretability.
- **Multiplicity** of good models.
- **Fairness** and **social** aspects.
- **Trust** of model producers and consumers.
- **Security** and **hacking.**
- **Regulated/controlled** environments **.**

# What's needed

## Building human-centered, low-risk models

# Building Human-Centered, Low-Risk Models

- Big picture.
- Interpretability focused.
- MLI module demo **only**.
- DAI auto ML models.*
- MLI UCs coverage by DAI.
- Techniques and algorithms.
- **Possible** workflow.
- IID and TS.
- Where MLI module fits in E2E.



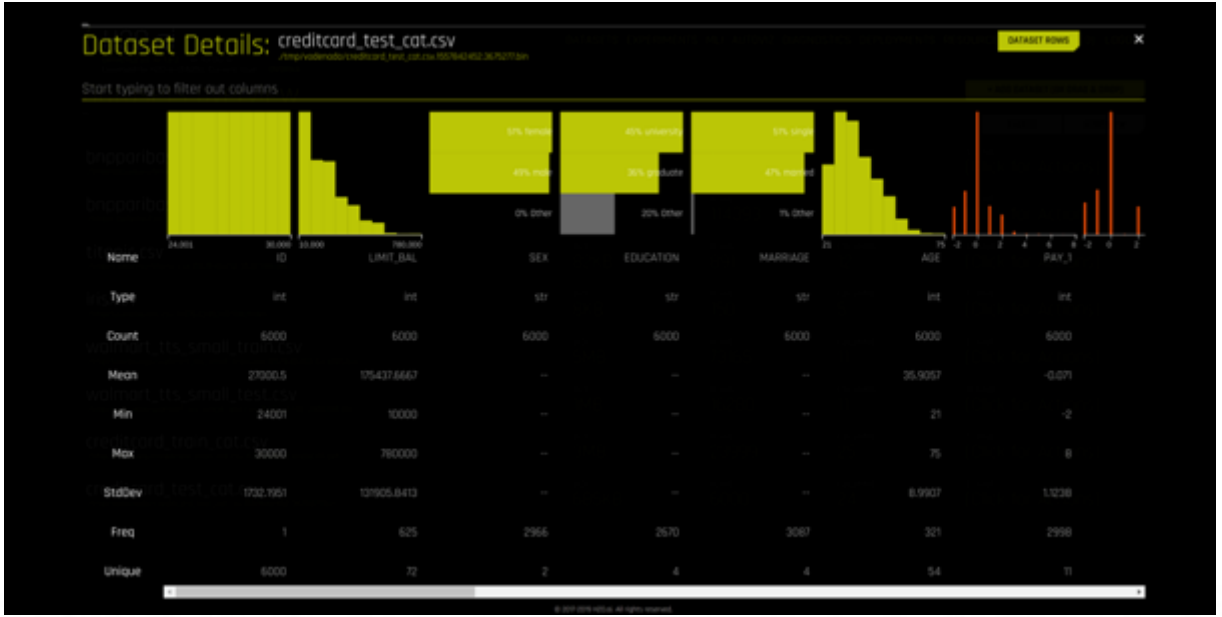

*) MLI module is not limited to DAI's models

# Building Human-Centered, Low-Risk Models

Load data

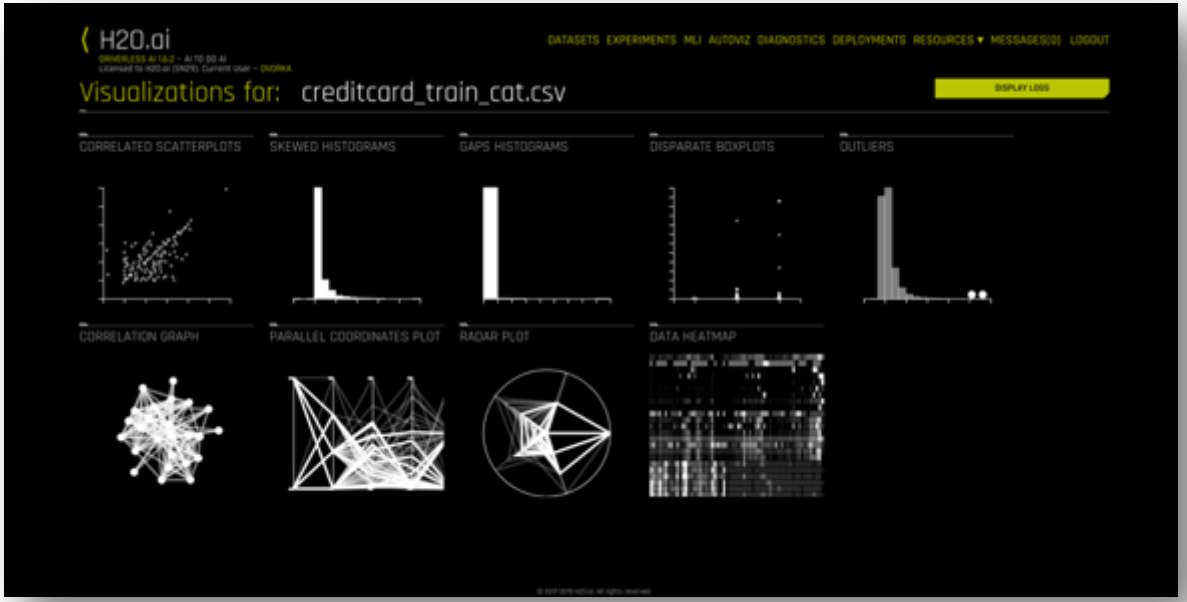
# Explanatory Data Analysis and Visualization

Load data → EDA



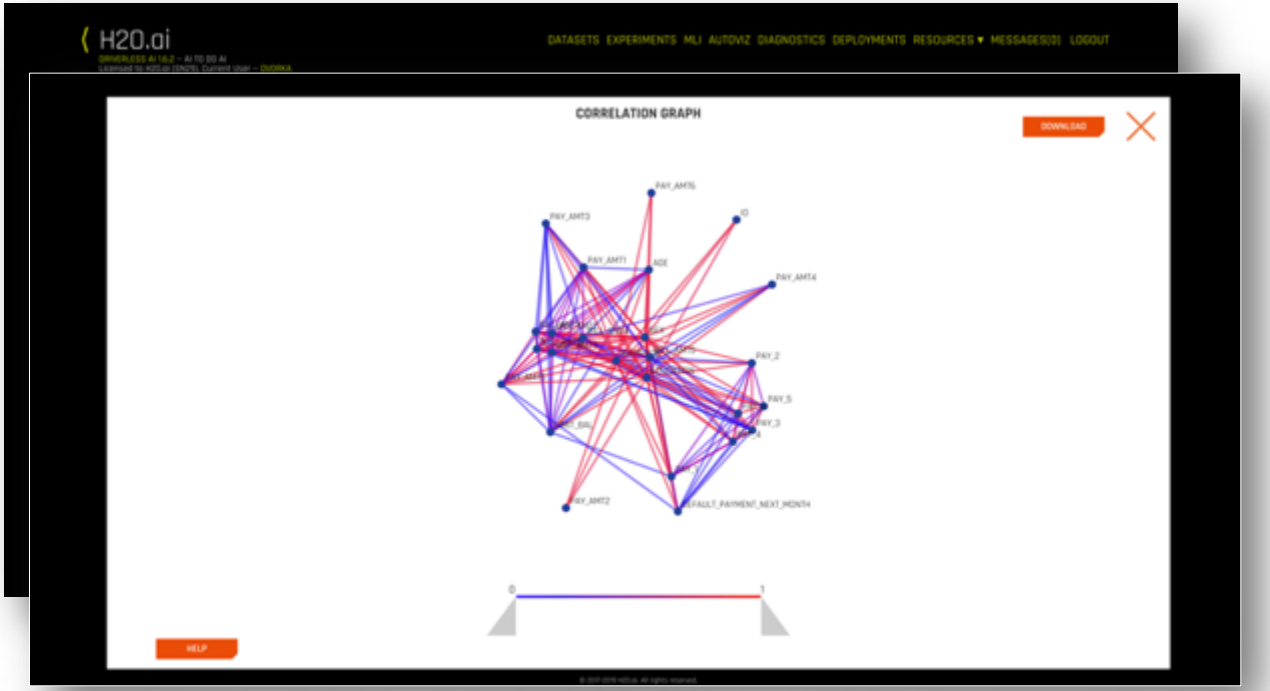
# Explanatory Data Analysis and Visualization

Load data → EDA



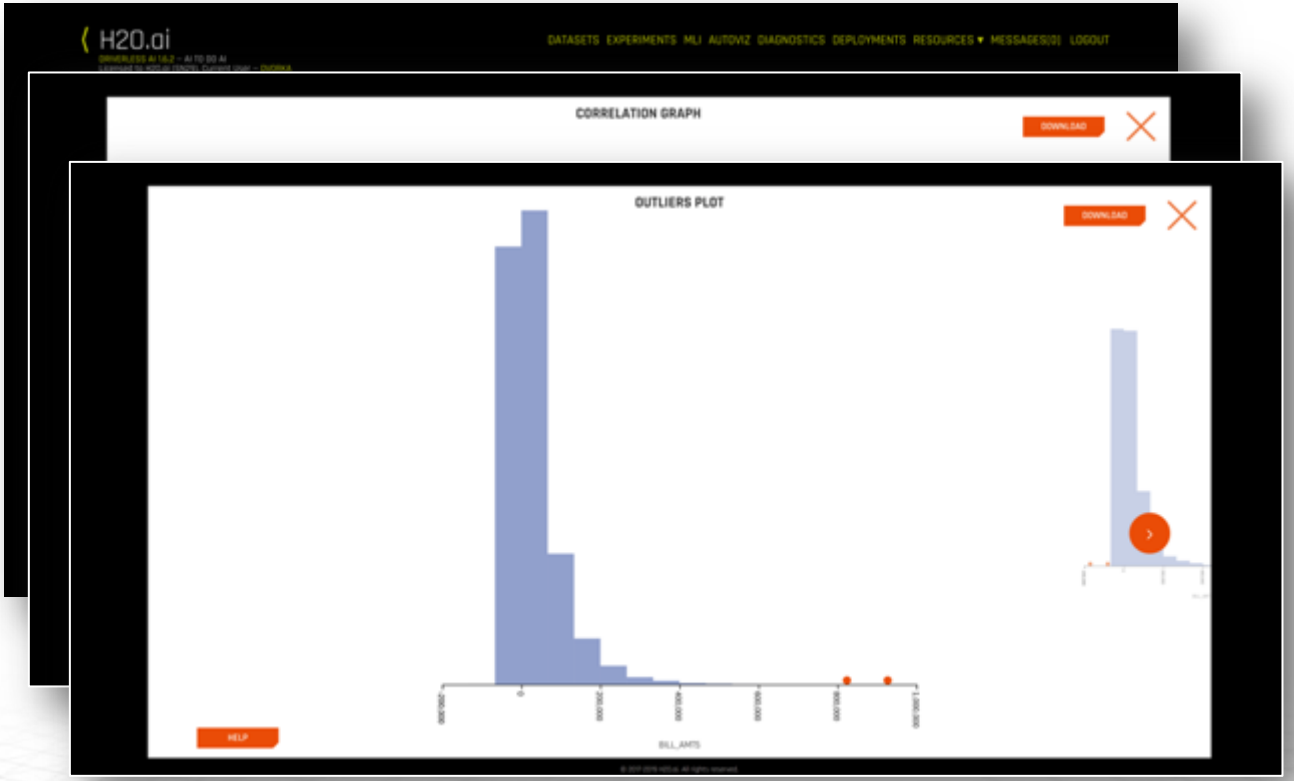
# Explanatory Data Analysis and Visualization

Load data → EDA



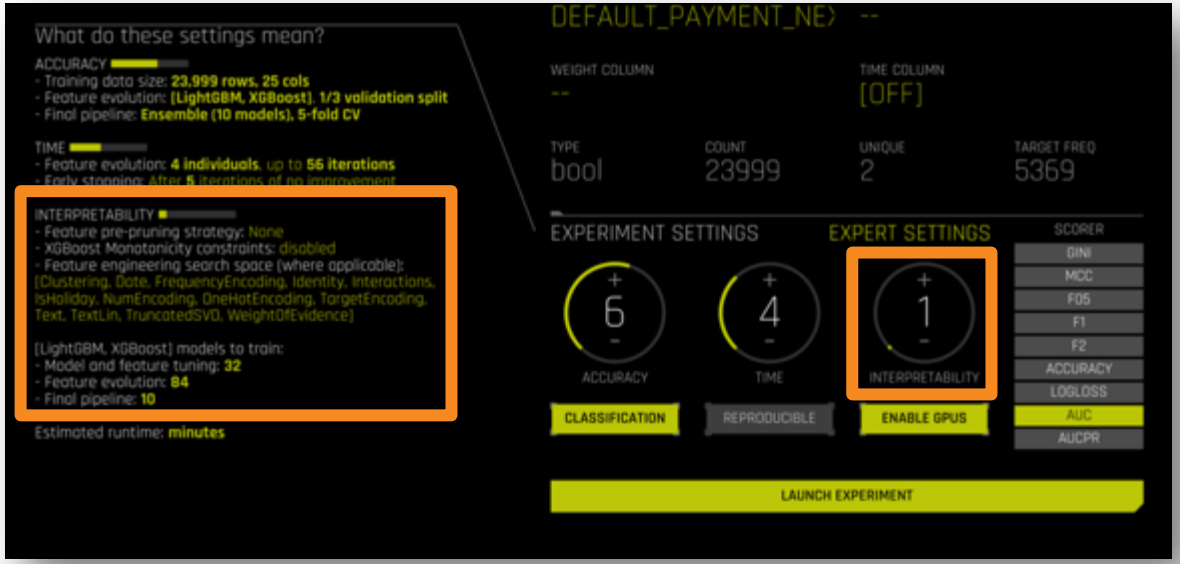
# Explanatory Data Analysis and Visualization

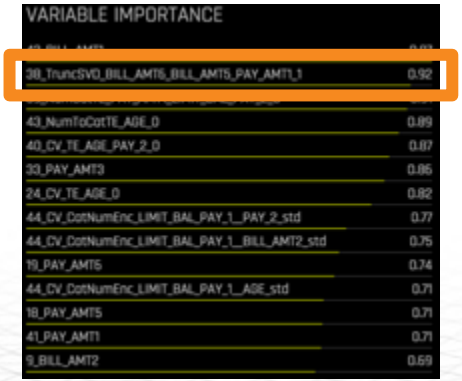
Load data → EDA



# Feature Engineering (Manual & Auto ML)

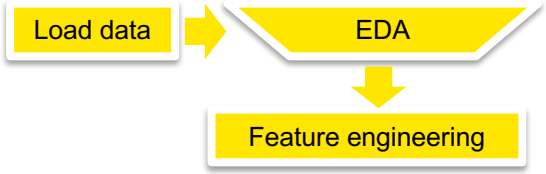
Load data → EDA → Feature engineering




Black box model

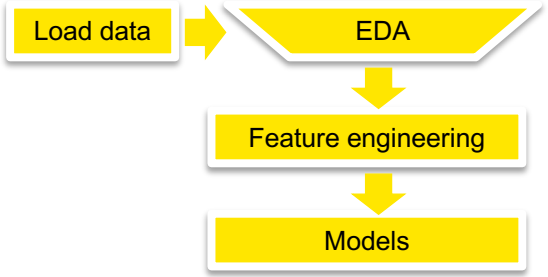
# Feature Engineering

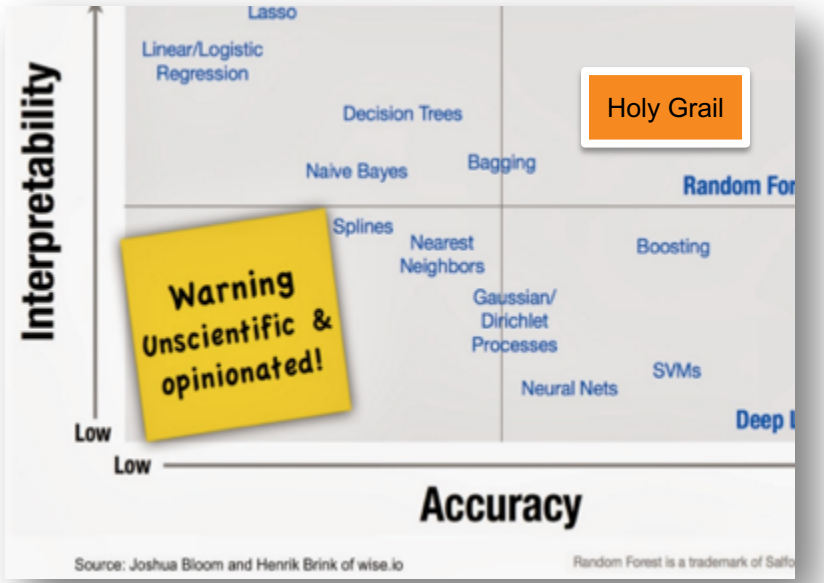




White box model

# Model Choice: Constrained, Simple, Fair

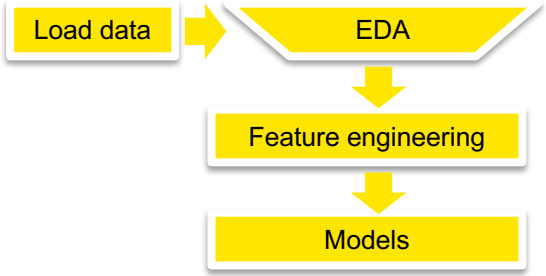


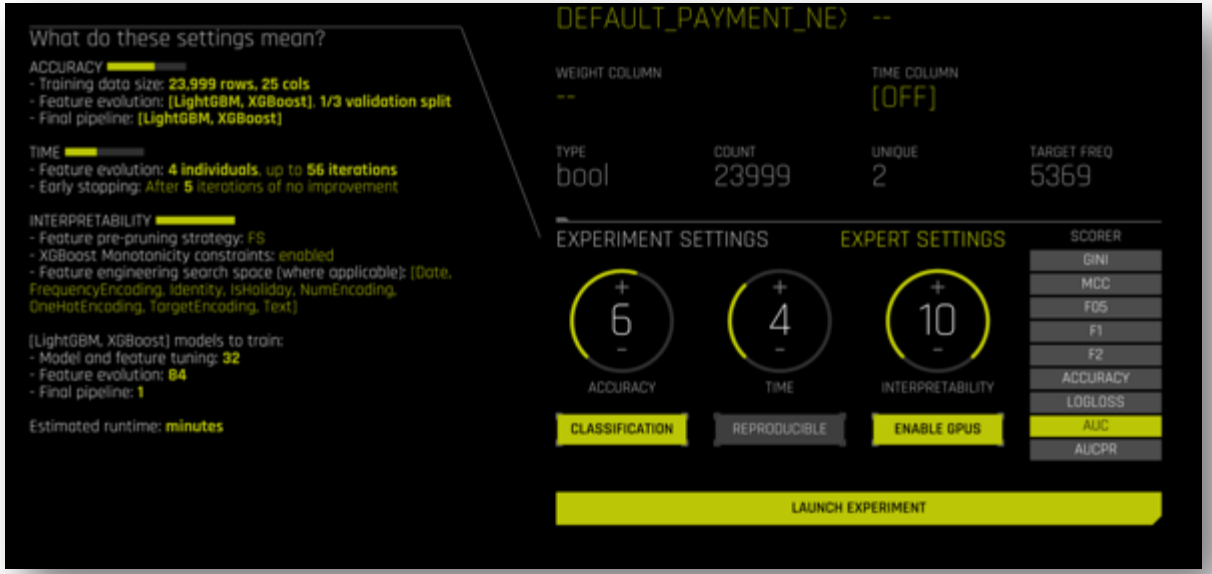




Black box model

White box model

# Model Choice: Constrained, Simple, Fair



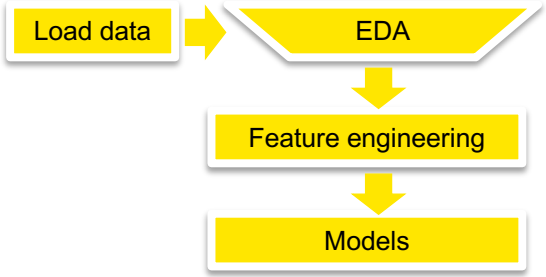




Black box model

White box model

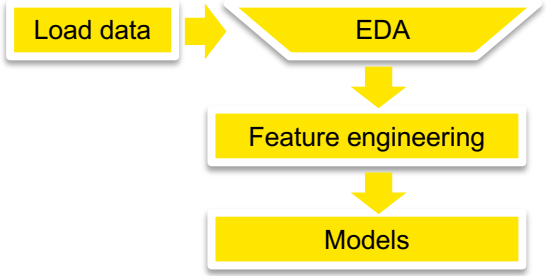
GLM (log regr.), Monotonic GBM (DT), XNN, …

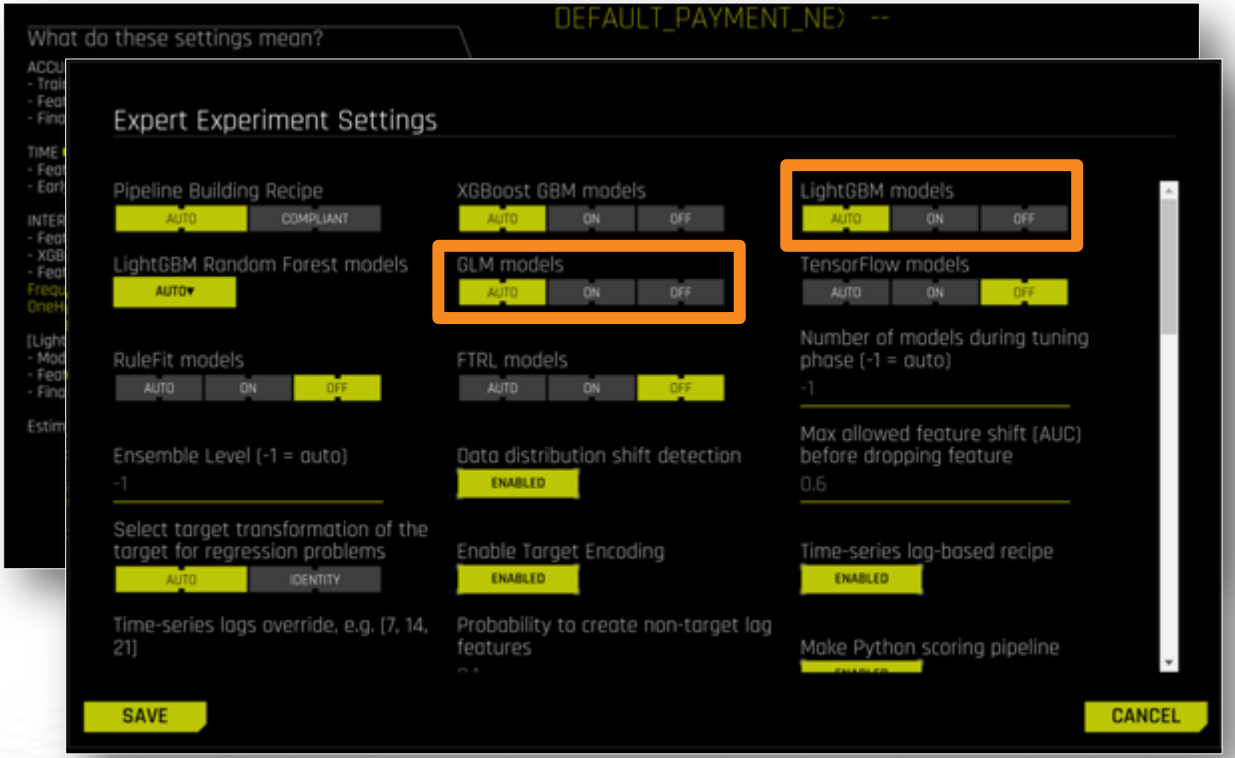
# Model Choice



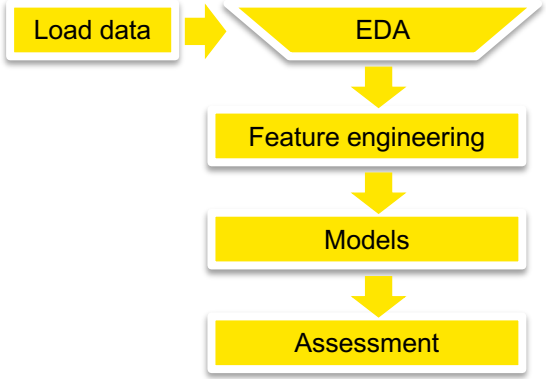

| Interpretability | Ensemble Level | Target Transformation | Feature Engineering | Feature Pre-Pruning | Monotonicity Constraints |
|---|---|---|---|---|---|
| 1 - 3 | <= 3 | | | None | Disabled |
| 4 | <= 3 | Inverse | | None | Disabled |
| 5 | <= 3 | Anscombe | Clustering (ID, distance)<br>Truncated SVD | None | Disabled |
| 6 | <= 2 | Logit<br>Sigmoid | | Feature selection | Disabled |
| 7 | <= 2 | | Frequency Encoding | Feature selection | Enabled |
| 8 | <= 1 | $4^{th}$ Root | | Feature selection | Enabled |
| 9 | <= 1 | Square<br>Square Root | Bulk Interactions (add, subtract, multiply, divide)<br>Weight of Evidence | Feature selection | Enabled |
| 10 | 0 | Identity<br>Unit Box<br>Log | Date Decompositions<br>Number Encoding<br>Target Encoding<br>Text (TF-IDF, Frequency) | Feature selection | Enabled |

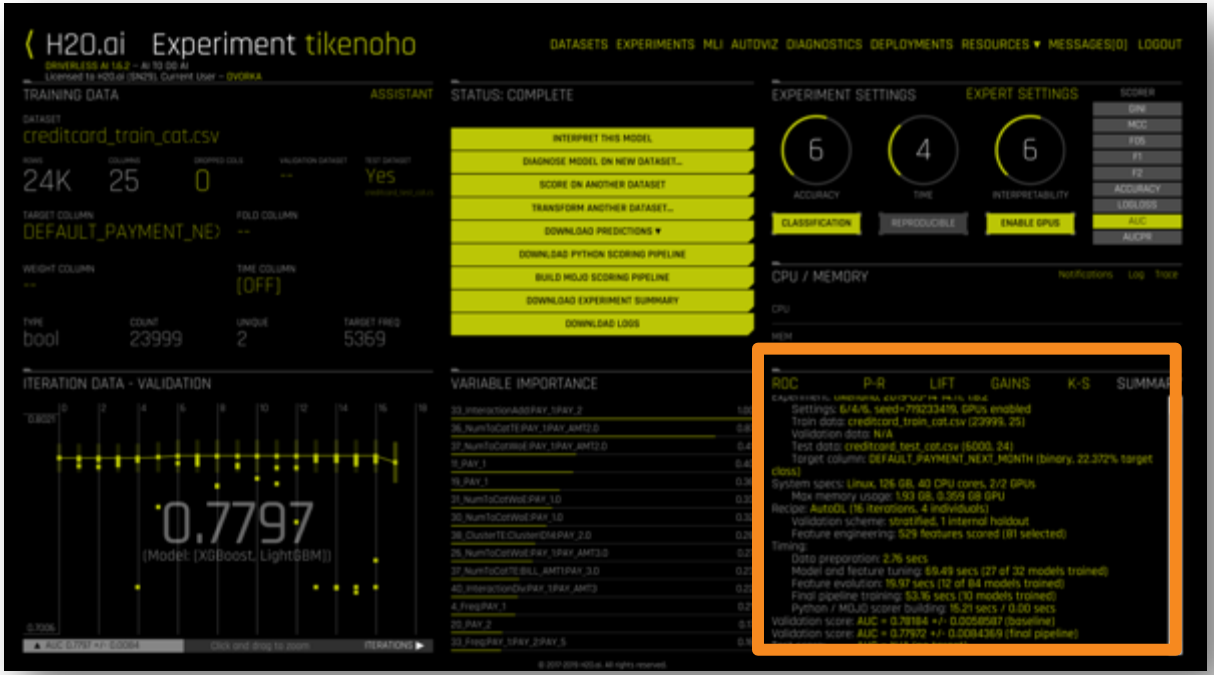
# Model Choice



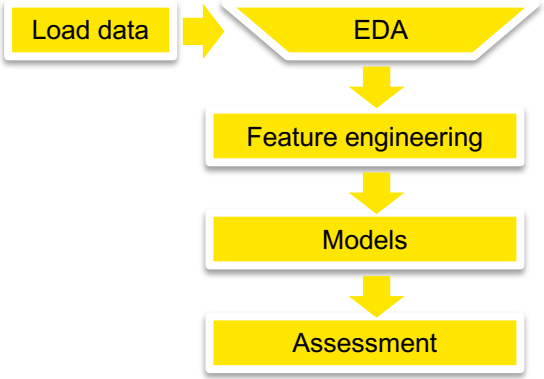



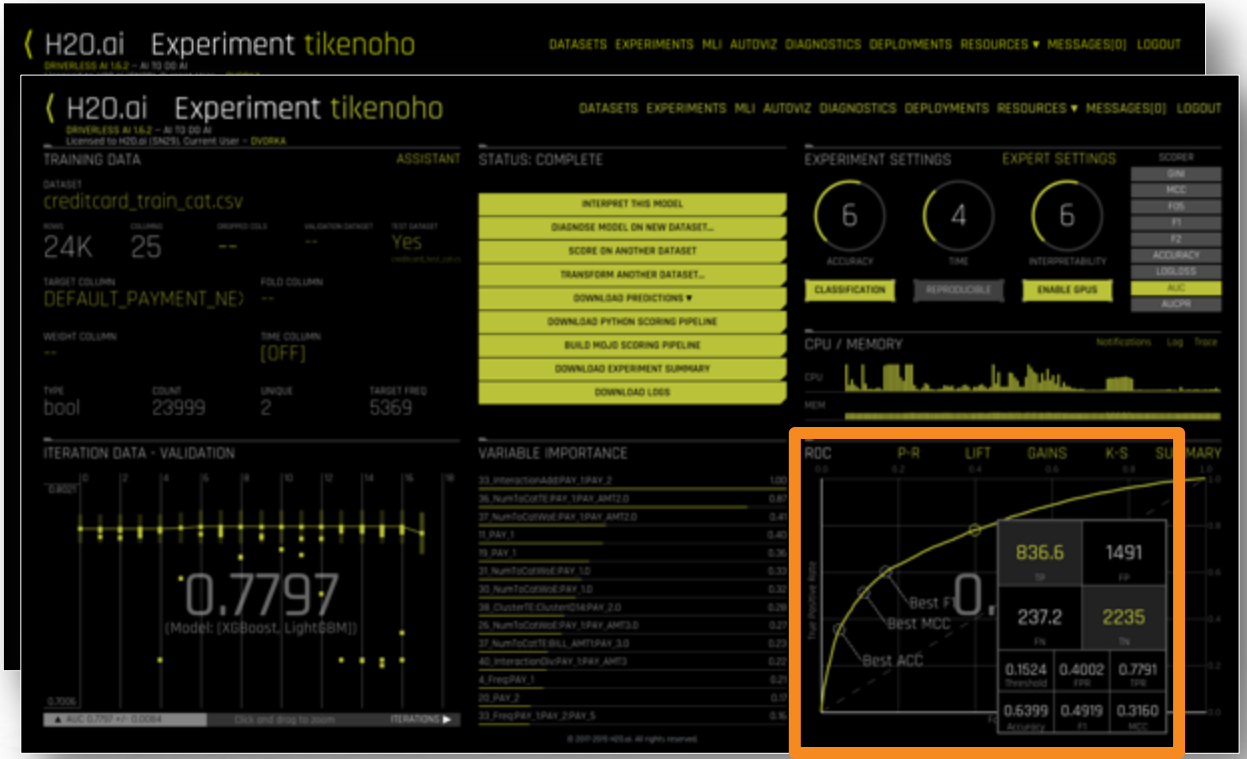
# Traditional Model Assessment and Diagnostics



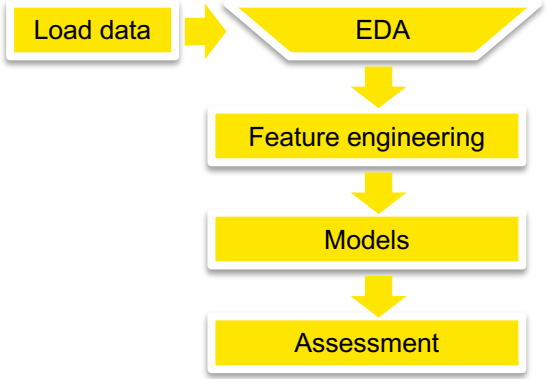



# Traditional Model Assessment and Diagnostics

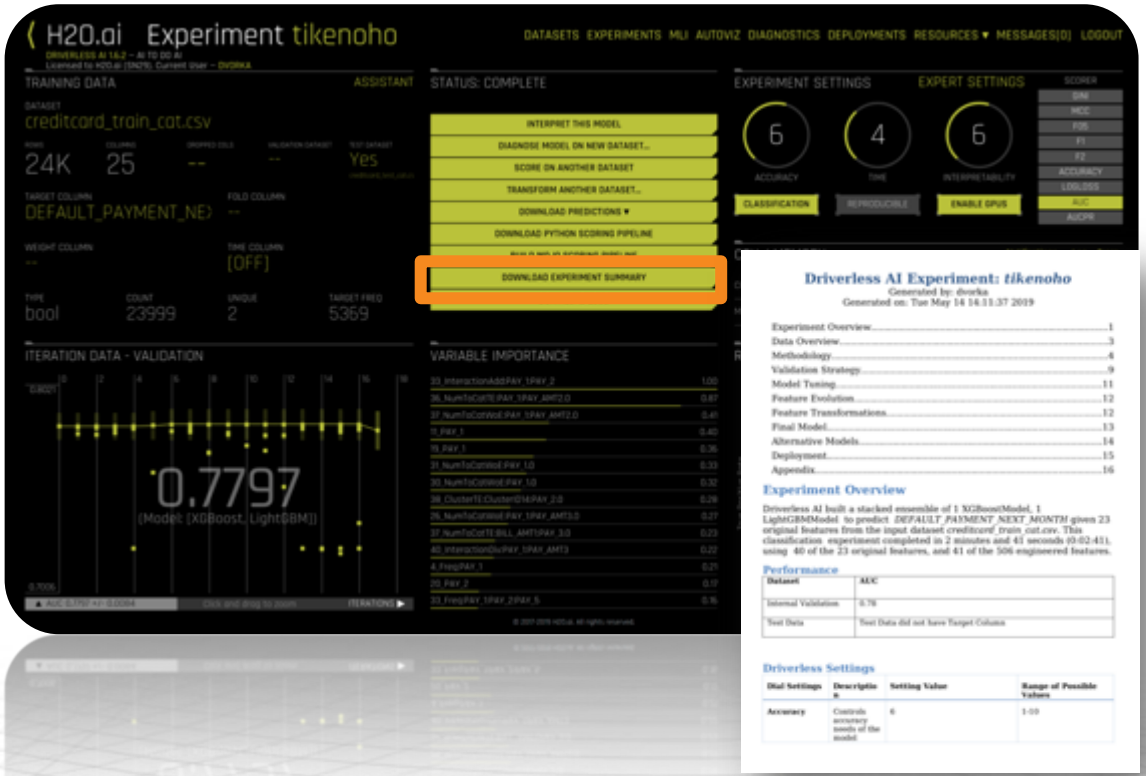


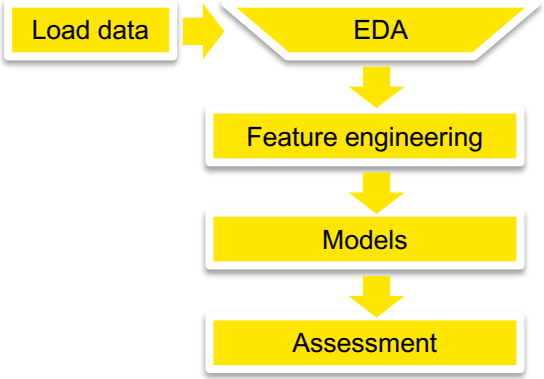
# Traditional Model Assessment and Diagnostics



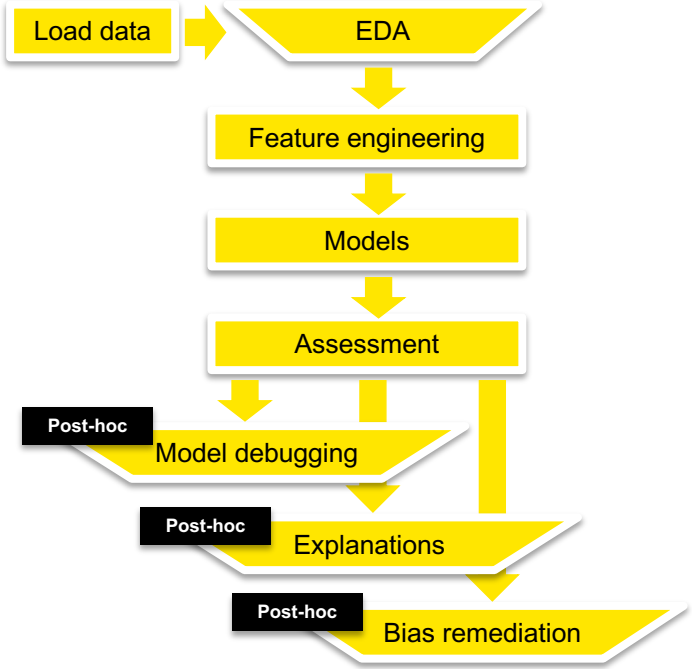

**Experiment summary** (document + YAML) + **AutoDoc**



# Traditional Model Assessment and Diagnostics

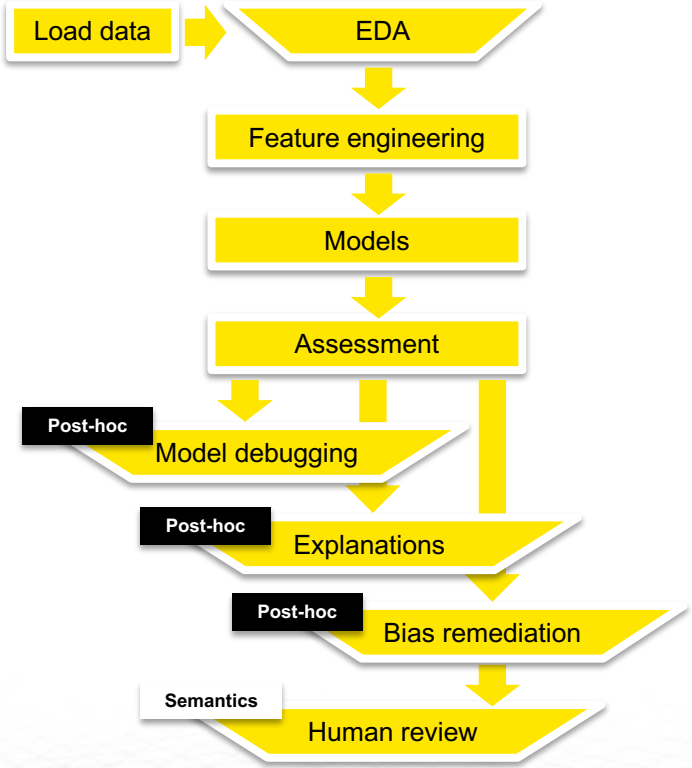


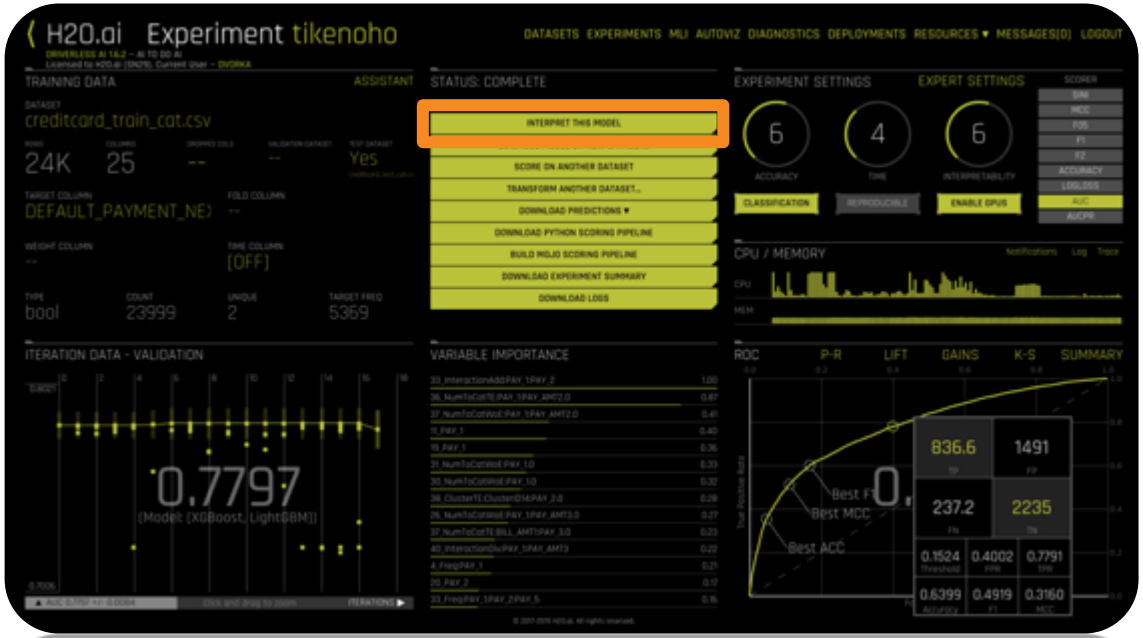

# Post-hoc Model Explanations and Debugging



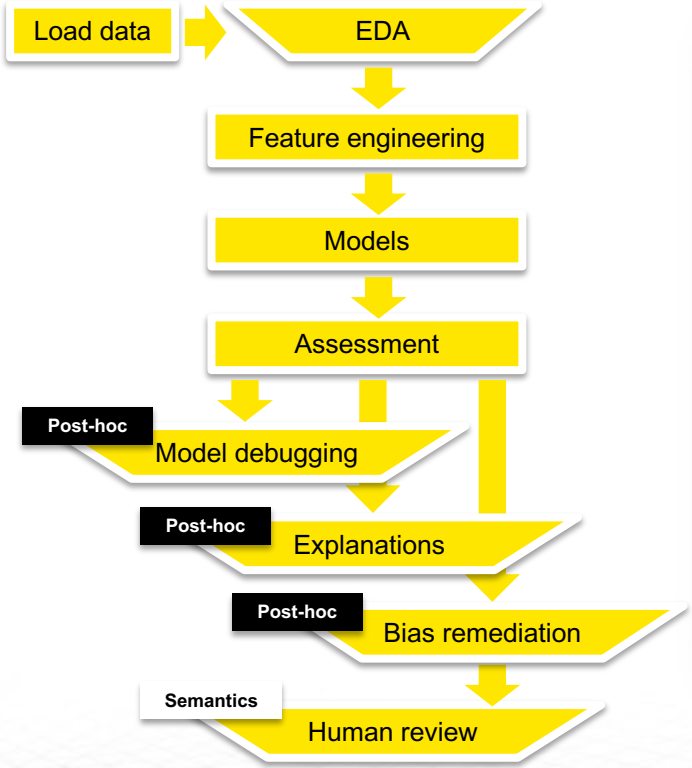

- **Post-hoc model debugging**
  - What-if, sensitivity analysis (accuracy).
- **Post-hoc explanations**
  - Reason codes.
- **Post-hoc bias assessment and remediation**
  - Disparate impact analysis.

# Human Review

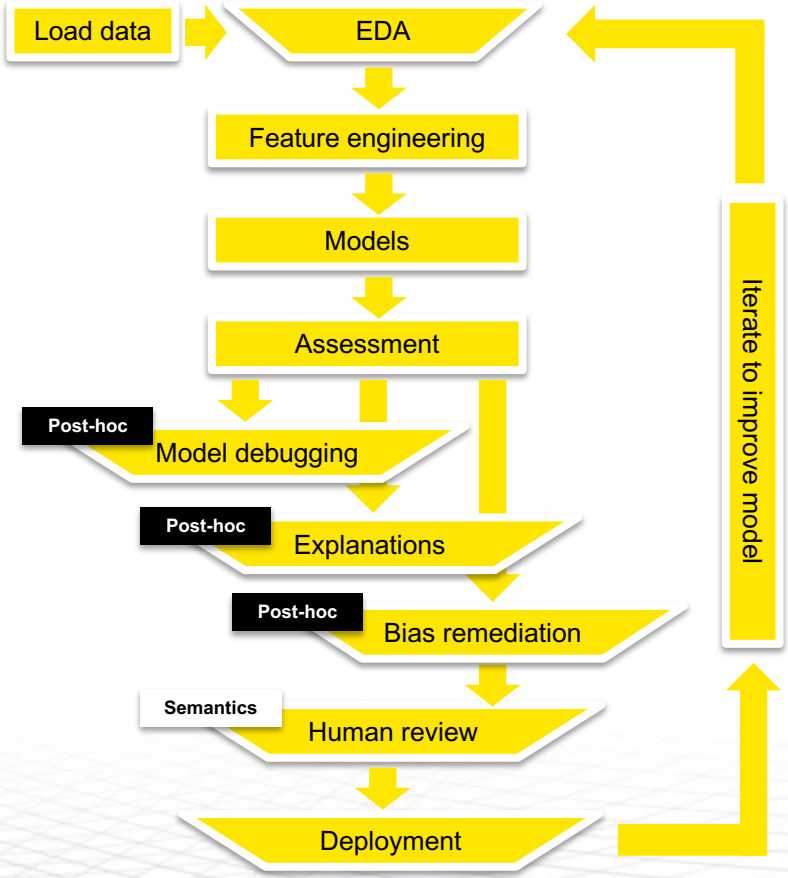


# Human Review




# Iterative Improvement



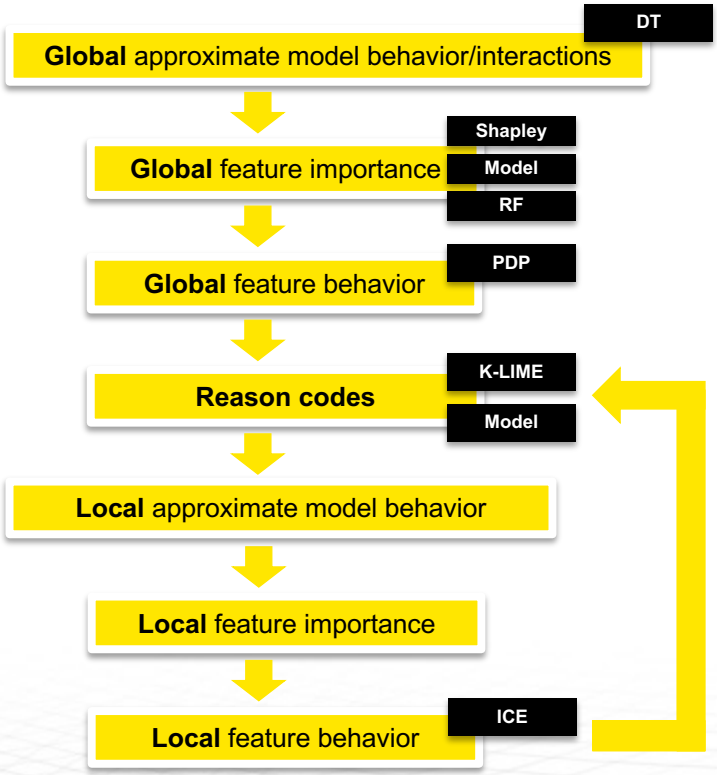
# How

*Explaining models - **MLI module** deep dive*

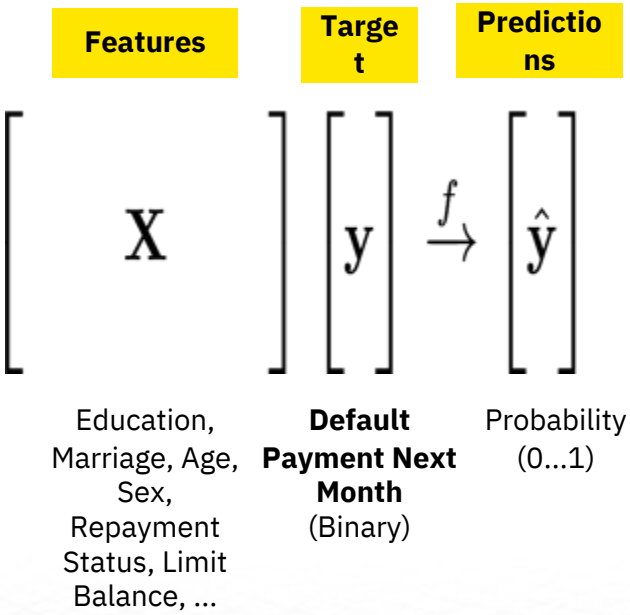
# H2O Driverless AI's **MLI module**

# IID and Time Series

# H2O Driverless AI's **MLI module**



# Demo Dataset: Credit Card (IID)



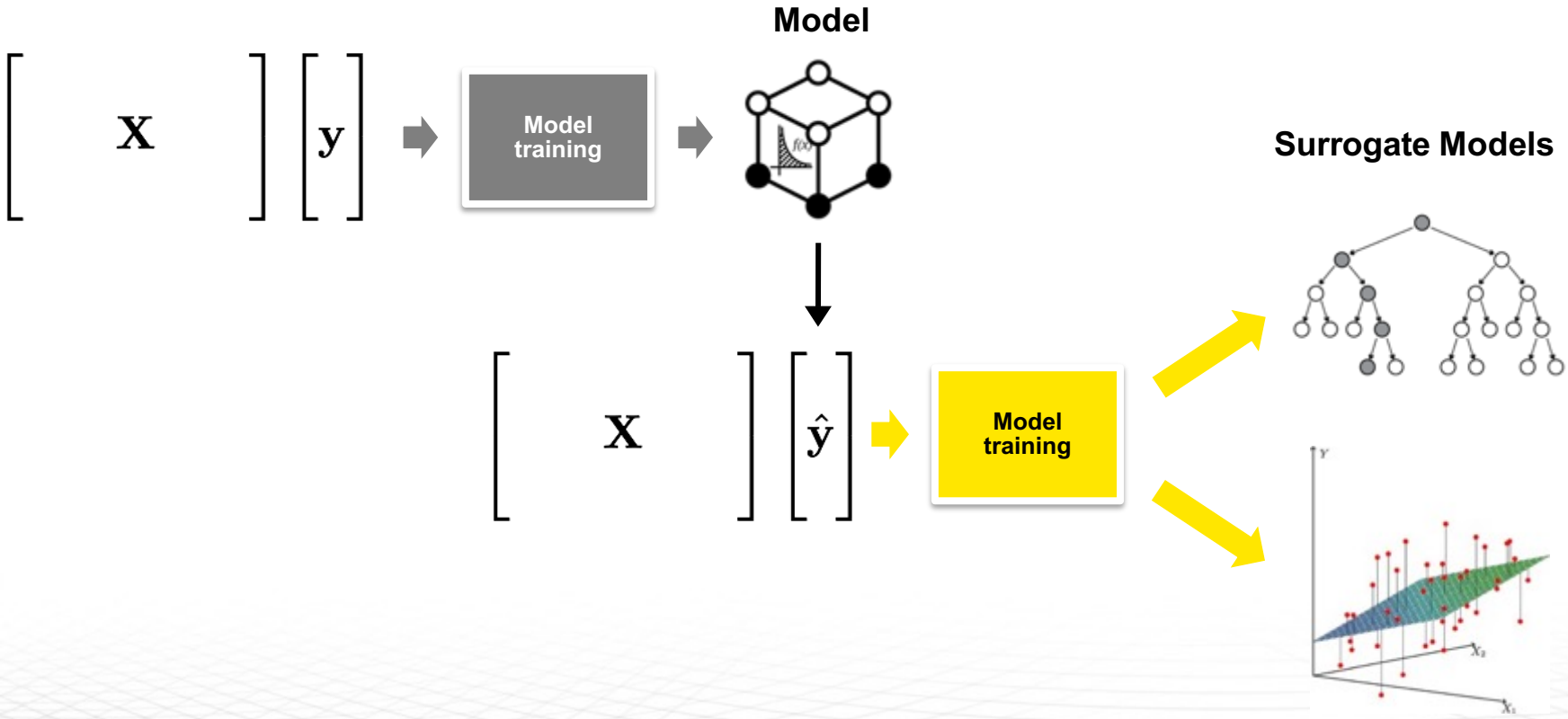

| Column Name | Description |
|---|---|
| ID | ID of each client |
| LIMIT_BAL | Amount of given credit in NT dollars (includes individual and family/supplementary credit) |
| SEX | Gender (1=male, 2=female) |
| EDUCATION | (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown) |
| MARRIAGE | Marital status (1=married, 2=single, 3=others) |
| AGE | Age in years |
| **PAY_x {1, …,6}** | Repayment status in August, 2005 – April, 2005 (-1=paid duly,1=payment delay for 1 month, …,8=payment delay for 8 months) |
| BILL_AMTx {1, …, 6} | Amount of bill statement in September, 2005 – April, 2005 (NT dollar) |
| PAY_AMTx {1, …, 6} | Amount of previous payment in September, 2005 – April, 2005 (NT dollar) |
| **default_payment_next_month** | Default payment (1=yes, 0=no) |

# Global Approximate Model Behavior/Interaction

- **Challenge:**
  - Black-box models
  - Original vs. transformed features.

- **Solution:** Surrogate models
  - **Pros**
    - Increases any black-box model's interpretability
    - Time complexity
  - **Cons**
    - Accuracy

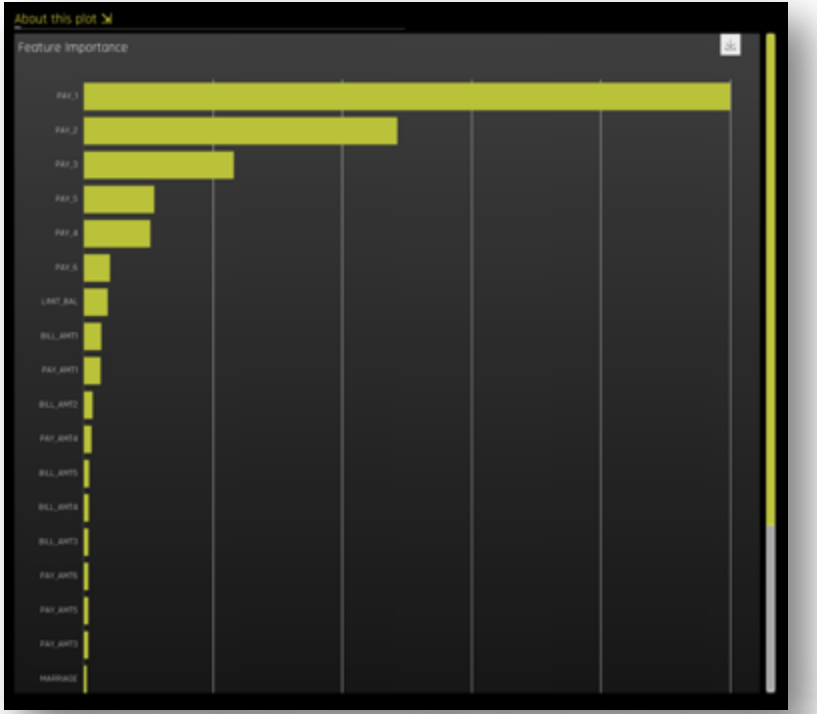
# Surrogate Models



# Global Feature Importance: Random Forest

- **Challenges:**
  - Black-box models
  - Original vs. transformed features

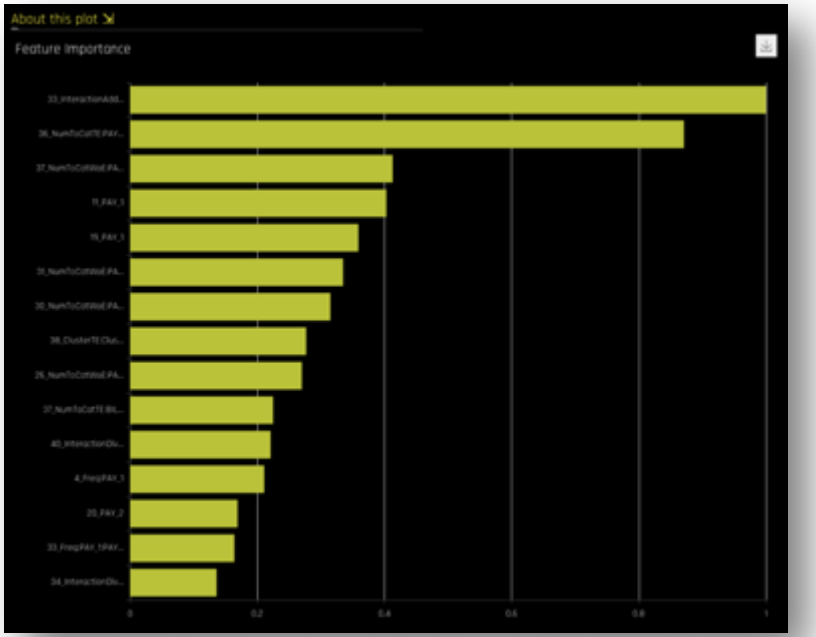
- **Solutions:**
  - Surrogate model: **RF** (introspection)
    - **Pros**:
      - Original features
      - Time complexity
    - **Cons**:
      - Accuracy



# Global Feature Importance: Original Model

- **Challenges:**
  - Black-box models
  - Original vs. transformed features

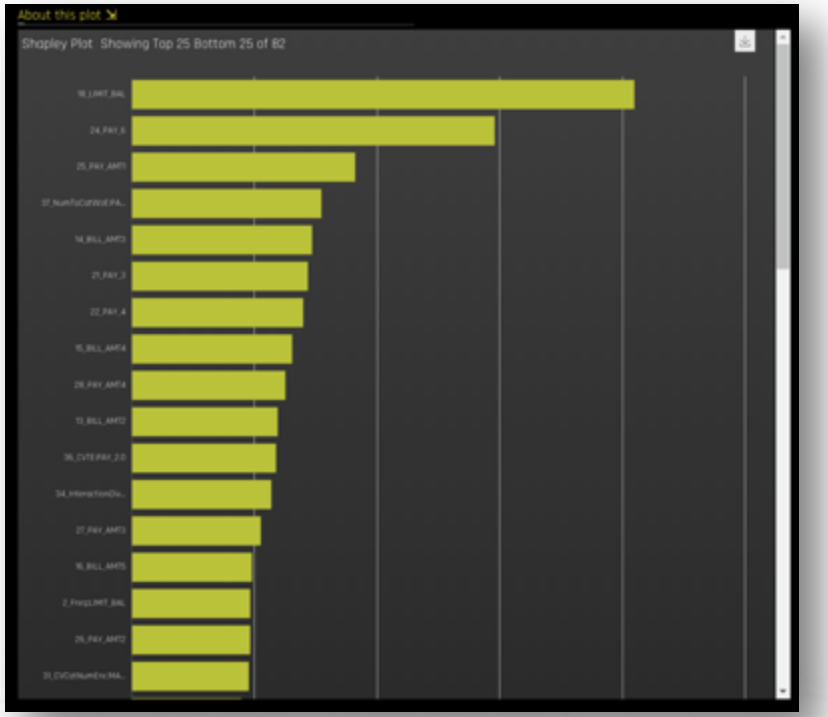
- **Solutions:**
  - **Original** (DAI) Model Introspection
    - **Pros**:
      - Accuracy
    - **Cons**:
      - Transformed features
      - Global only



# Global Feature Importance: Shapley Values

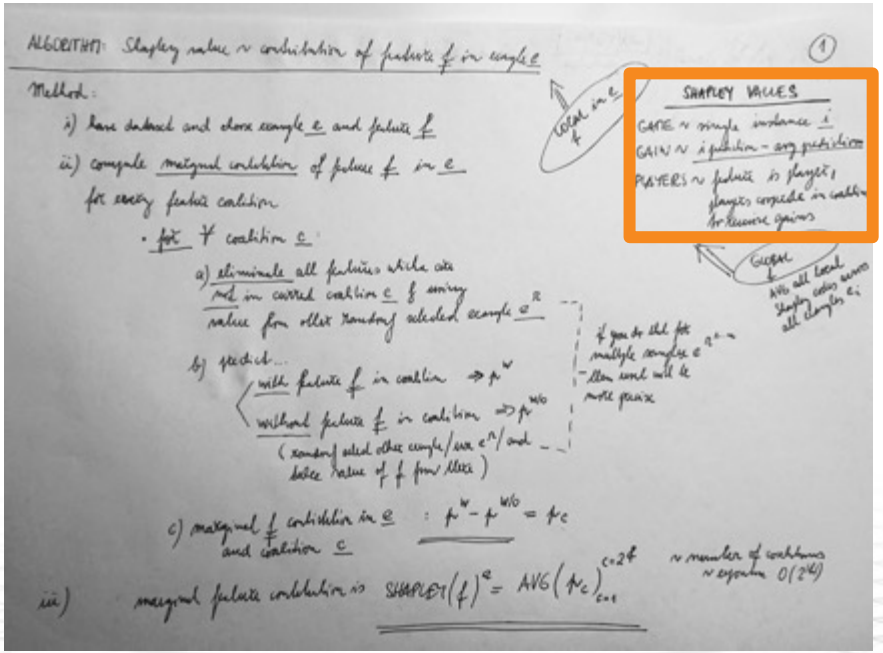
- **Challenge**
  - Black-box models
  - Original vs. transformed features

- **Solutions:**
  - Shapley values
    - Pros:
      - Accuracy
      - Math correctness
    - Cons:
      - Time complexity
      - Transformed features



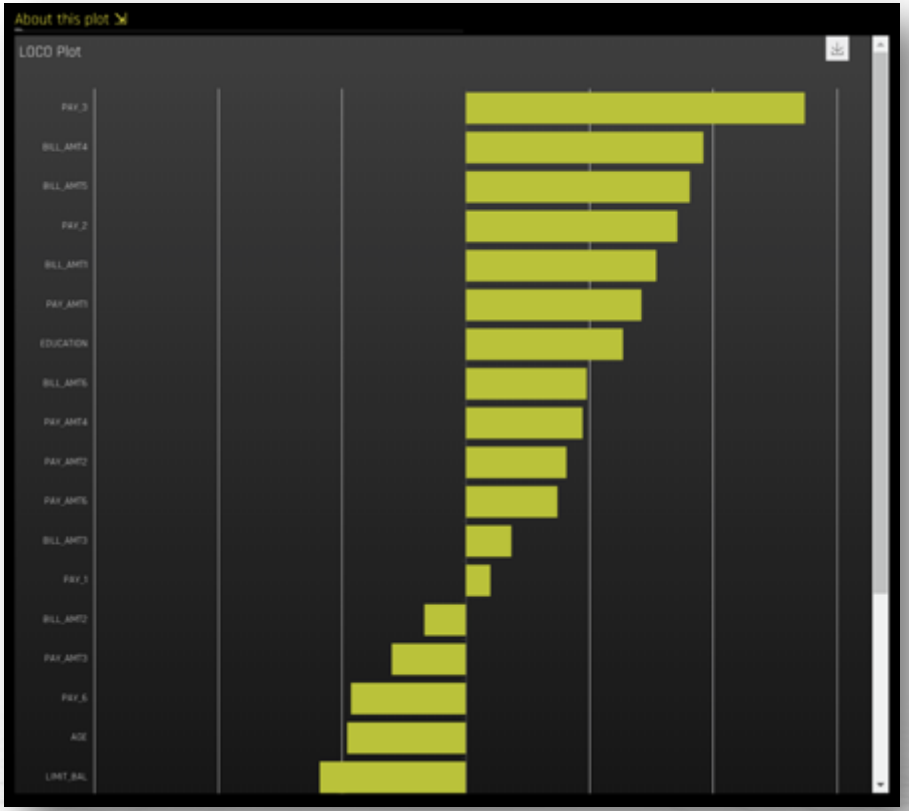
# Shapley Values

- Lloyd Shapley
  - Americal mathematician who won **Nobel** prize in 2012 (Economics).
  - Shapley values was his Ph.D. thesis written in **50s**.
- **Shapley values:**
  - Supported by **solid** mathematical (game) theory.
  - Calculation has **exponential** time complexity (number of coalitions) .
  - Typically **unrealistic to compute** in real world.
  - Can be computed in **global** or **local** scope.
  - **Guarantee fair distribution** among features in the instance.
  - Does **not** work well in **sparse** cases, **all** features must be used.
  - Return **single value per feature**, not a model.



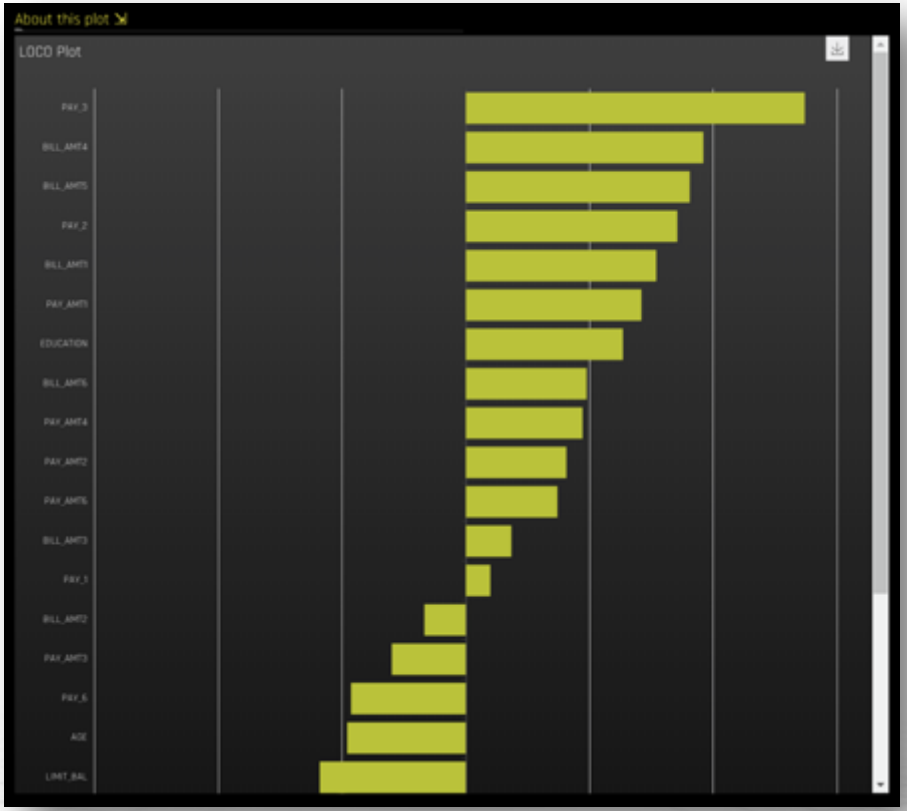
# Feature importance: Leave One Covariate Out

- **UC:**
  - Complete other feature importance charts with bias tendency
- **Challenge:**
  - Black-box models
- **Solution:**
  - LOCO



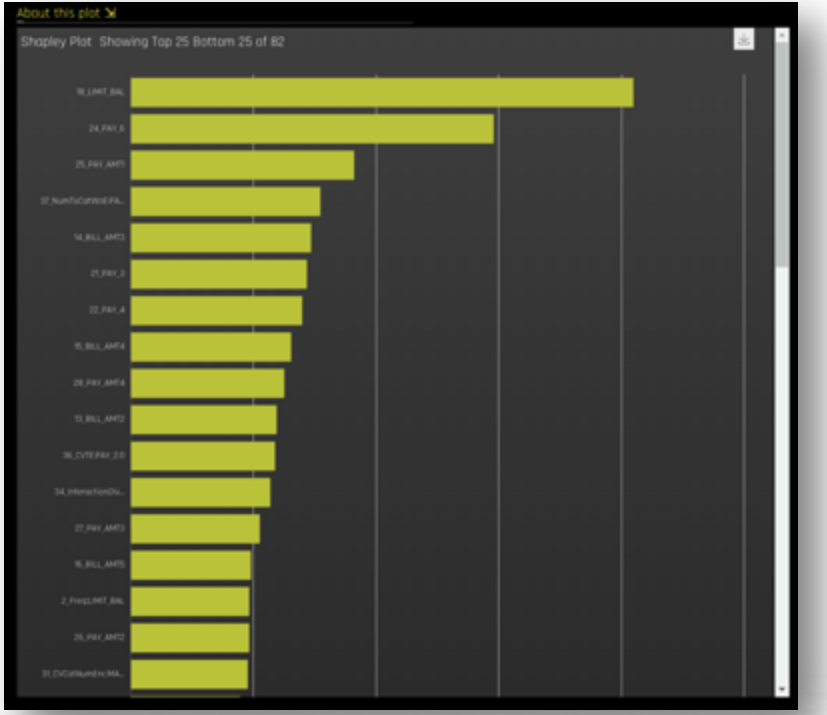
# Feature importance: Leave One Covariate Out

| Sex | Age | ... | Fare | ŷ | ŷ(-Sex) | ŷ(-Age) | ... | ŷ(-Fare) |
|---|---|---|---|---|---|---|---|---|
| M | 11 | ... | 8.45 | 0.2 | 0.01 | 0.1 | ... | 0.21 |
| F | 34 | ... | 51.86 | 0.8 | 0.6 | 0.65 | ... | 0.78 |
| M | 26 | ... | 21.08 | 0.5 | 0.2 | 0.3 | ... | 0.53 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |



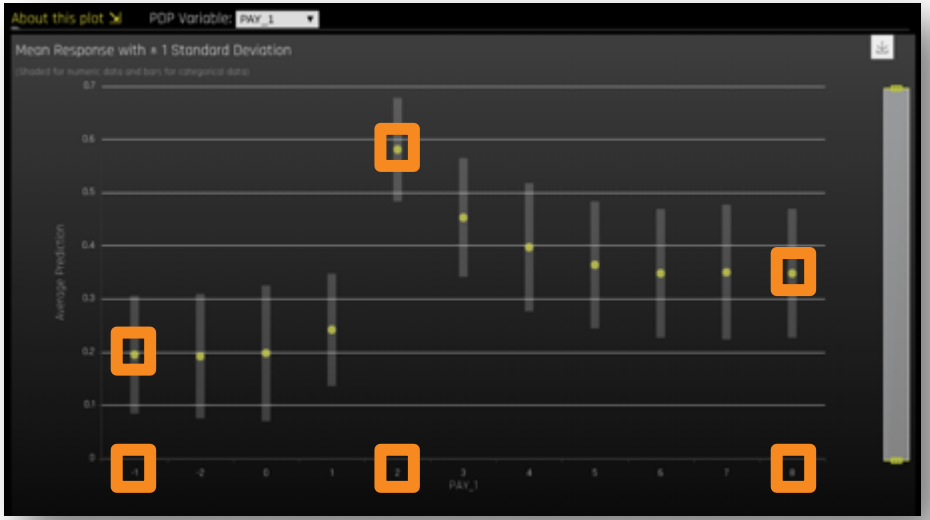
# Global Feature Importance

- **Methods**
  - Surrogate models:
    - RF (introspection)
    - Leave One Covariate Out (LOCO)
  - Original model (introspection)
  - **Shapley values**



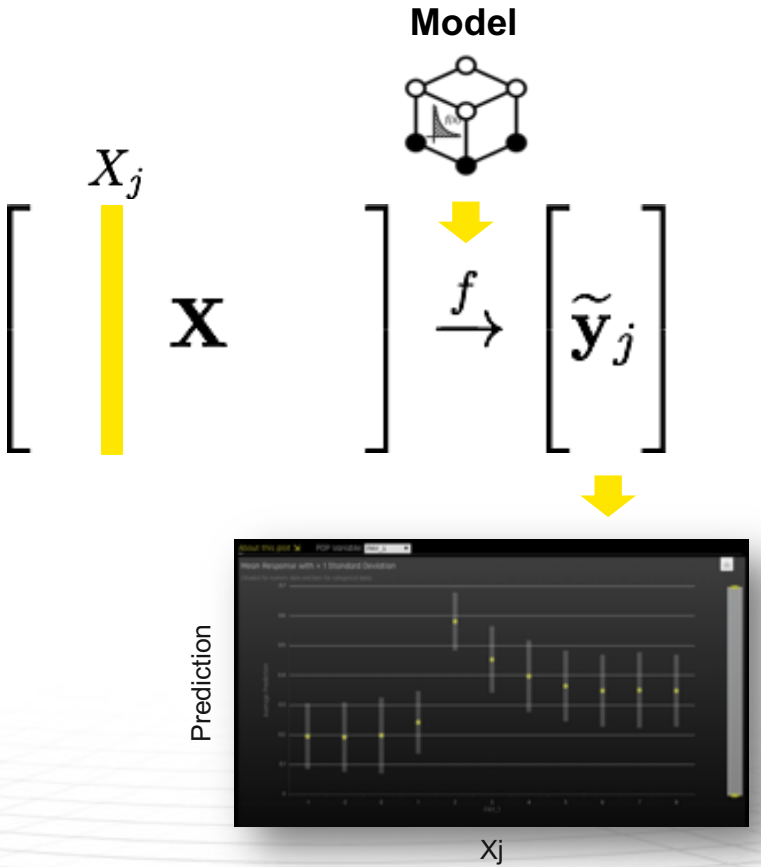
# Global Feature Behavior: Partial Dependence Plot

- **Solution:** Surrogate model PDP
  - **Pros**
    - Time complexity
    - Original features
    - White/black model interpretability
  - **Cons**
    - Accuracy

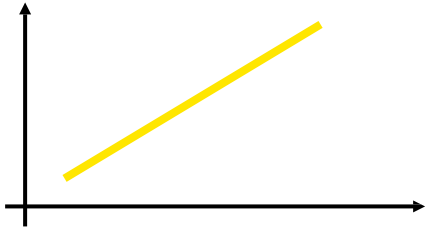


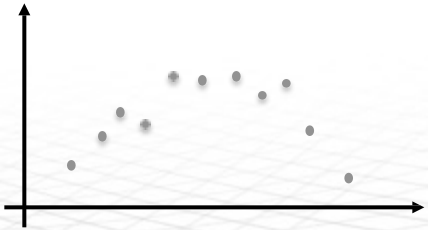
# Global Feature Behavior: Partial Dependence Plot

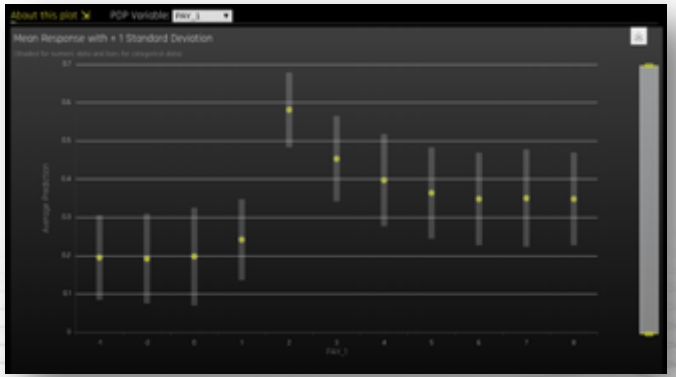
- **Solution:** Surrogate model PDP
  - **Pros**
    - Time complexity
    - Original features
    - White/black model interpretability
  - **Cons**
    - Accuracy



# PDP: Character of the Feature Behavior





# Reason codes: Local Feature Importance

**Global** approximate model behavior/interactions — DT

↓

**Global** feature importance — Shapley, Model, RF

↓

**Global** feature behavior — PDP

↓

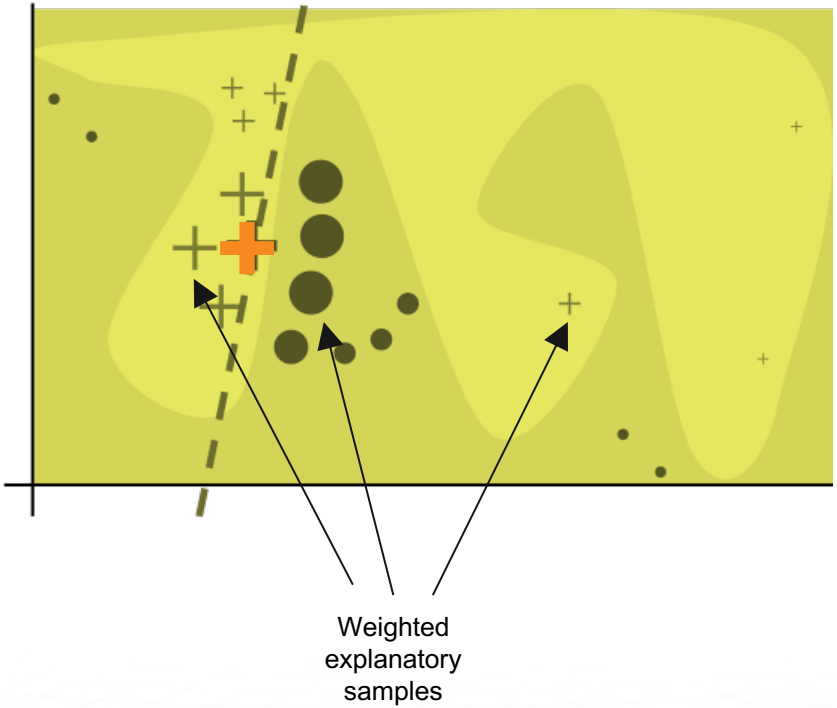
**Reason codes** — K-LIME, Model

# Reason codes: Local Feature Importance

- **UCs:**
  - Predictions explanations
  - Legal
  - Debugging
  - Drill-down,
  - …
- From **global to local** scope
- Surrogate methods:
  - **K-LIME** (K-means)
  - LIME-SUP (trees)

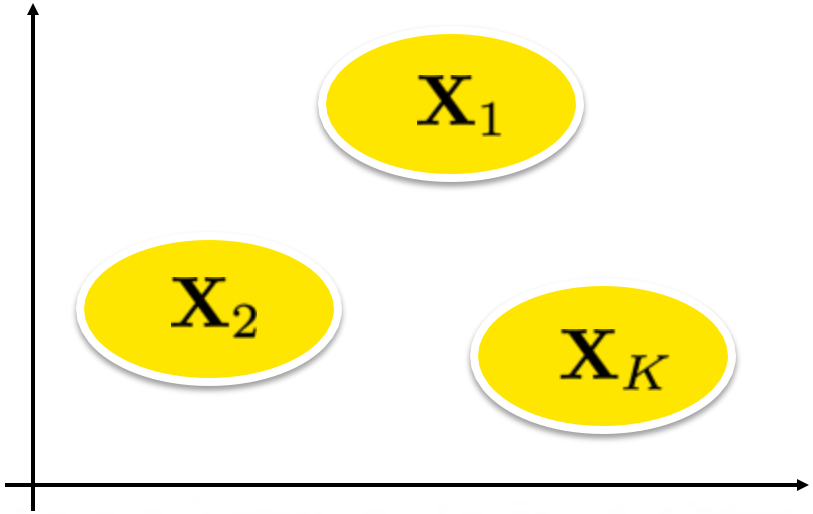
# LIME: Local Interpretable Model-agnostic Explanations

- Weighted **linear** surrogate model used to explain **non-linear** decision boundary in **local region.**
- **Single prediction.**
- ✚ example:
  - Set of explainable records are scored using the **original model**.
  - To interpret a decision about another record, the explanatory records are **weighted by their closeness** to that record.
  - L1 regularized linear model is trained on this weighted explanatory set.
  - The **parameters of the linear model** then help **explain** the prediction for the selected record.

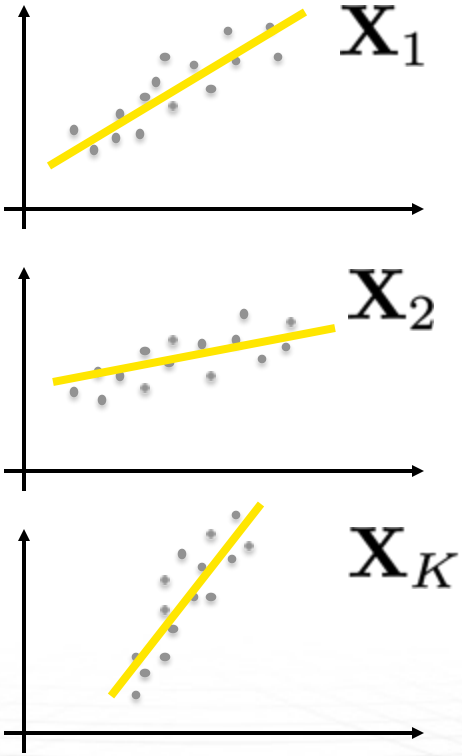



Source: https://github.com/marcotcr/lime

# K-LIME: Clustered LIME




$$\{\mathbf{X}_1 \cup \mathbf{X}_1 \cup \ldots \mathbf{X}_K\} = \mathbf{X}$$



# Reason codes: Local Feature Importance

- **UCs:**
  - Predictions explanations
  - Legal
  - Debugging
  - Drill-down,
  - …

- From **global to local** scope
- From global explanatory model to **cluster-scoped** explanatory model.

# Reason codes: Local Feature Importance

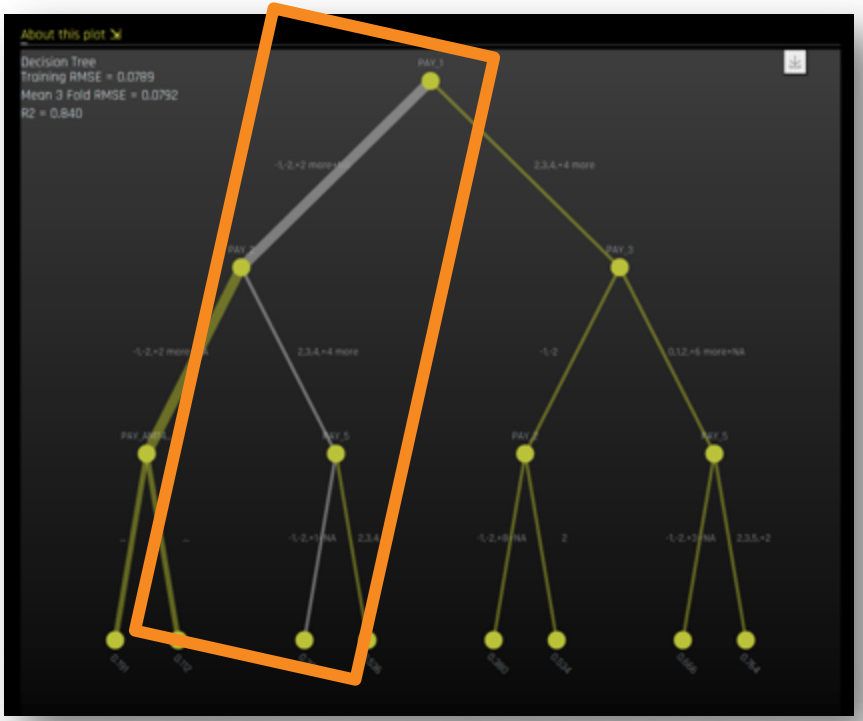
- **Challenges:**
  - Black-box models
  - Original vs. transformed features
- **Solutions:**
  - Surrogate model: **K-LIME**
    - **Pros**:
      - Original features
      - Time complexity
    - **Cons**:
      - Accuracy


$$g(\mathbf{x}^{(i)}) \approx h_{\mathrm{GLM},k}(\mathbf{x}^{(i)}) = \beta_0^{[k]} + \sum_{j=1}^{P} \beta_j^{[k]} x_j^{(i)}$$

reason code

# Local Approximate Model Behavior/Interaction

- **UC:**
  - Particular instance explanation
  - Note path segments thickness.
- **Challenge:**
  - Black-box models
- **Solution:** Surrogate models
  - **Pros**
    - Black-box models interpretability
    - Time complexity
  - **Cons**
    - Accuracy

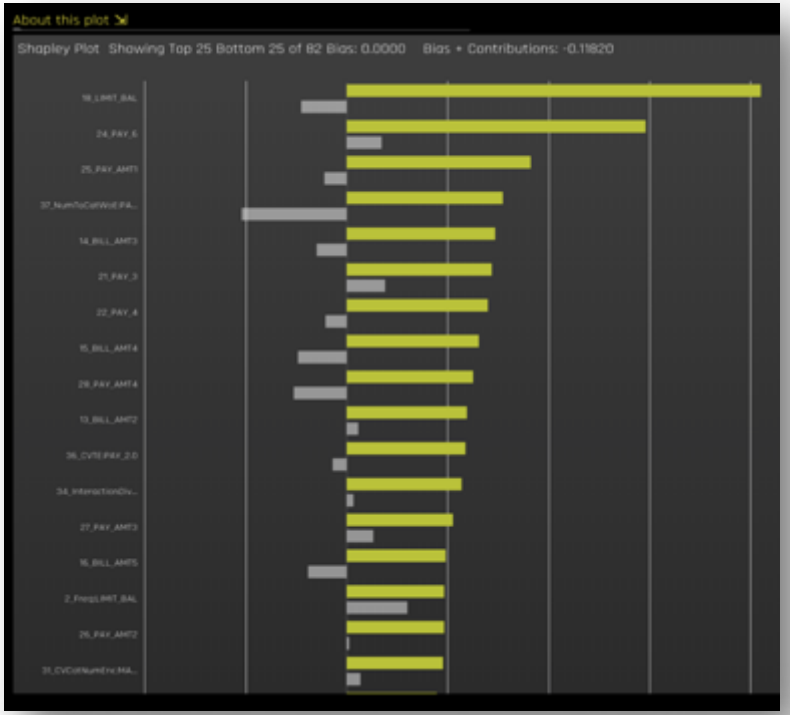


# Local Feature Importance

- Mean absolute value vs. **local contributions**

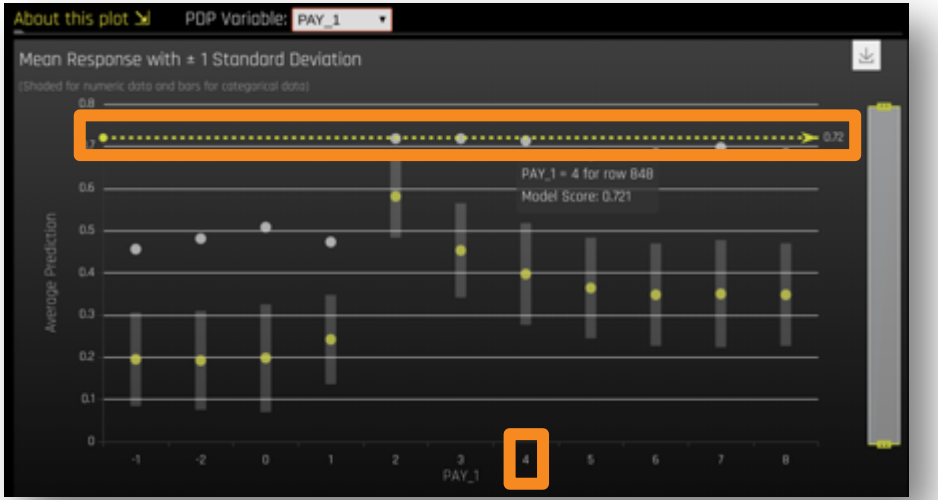
- **Challenge**
  - Black-box models
  - Original vs. transformed features

- **Solutions:**
  - Surrogate models:
    - RF (introspection)
    - Leave One Covariate Out (LOCO)
  - **Shapley values**



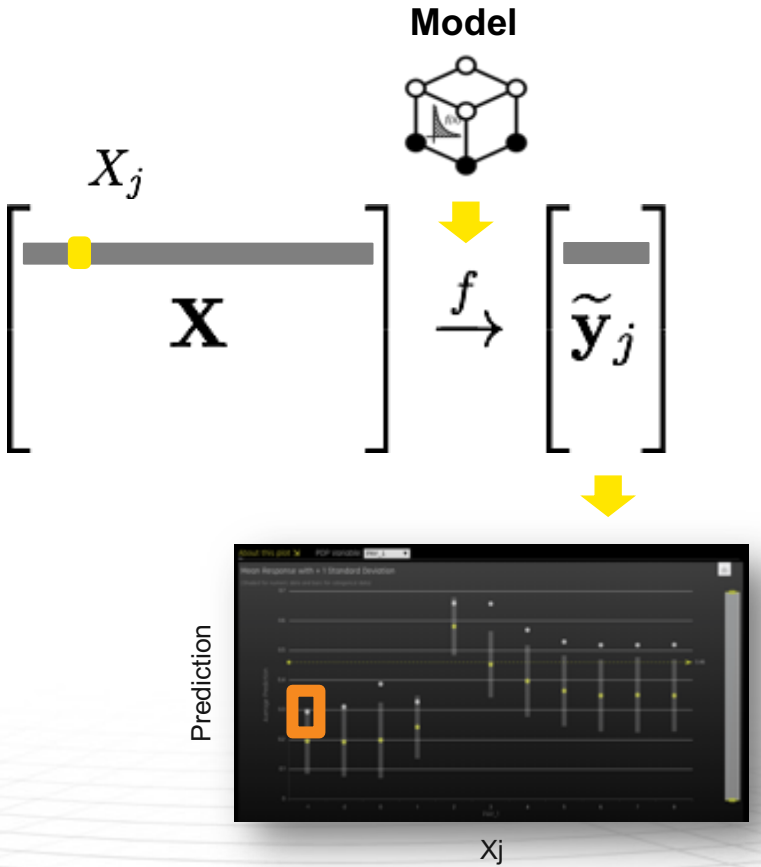
# Local Feature Behavior: ICE

- **Solution:** Surrogate model ICE
  - **Pros**
    - Time complexity
    - Original features
    - White/black model interpretability
  - **Cons**
    - Accuracy
      (dotted line vs. gray dot discrepancy)



# ICE: Individual Conditional Expectations

- **Solution:** Surrogate model ICE
  - **Pros**
    - Time complexity
    - Original features
    - White/black model interpretability
  - **Cons**
    - Accuracy



# AutoDoc

## Driverless AI Experiment: *cuwecoga*

Generated by: username
Generated on: Wed May 15 12:04:26 2019



### Experiment Overview

Driverless AI built 1 LightGBMModel to predict *DEFAULT_PAYMENT_NEXT_MONTH* given 23 original features from the input dataset *CreditCard_Cat-train.csv*. This classification experiment completed in 1 minutes and 19 seconds (0:01:19), using 20 of the 23 original features, and 4 of the 90 engineered features.

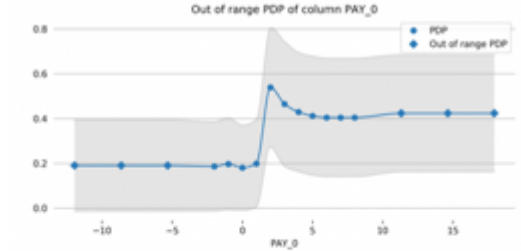
### Performance

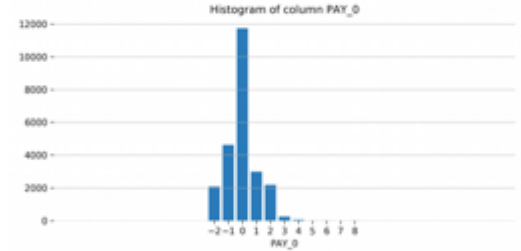
| Dataset | AUC |
|---|---|
| Provided Validation Data | 0.771 |
| Test Data | Test Data not Provided |

### Driverless Settings

| Dial Settings | Description | Setting Value | Range of Possible Values |
|---|---|---|---|
| Accuracy | Controls accuracy | 3 | 1-10 |

Feature **PAY_0**

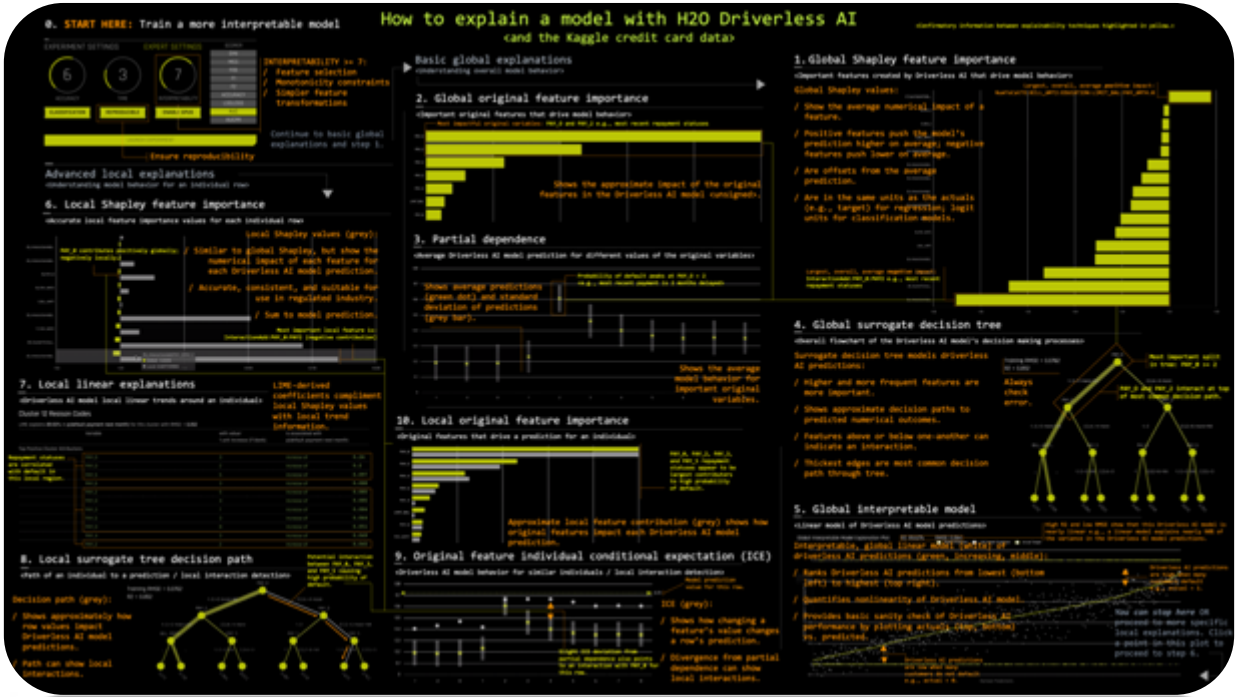


# MLI for **Time Series**

- Time series experiments
  - Test dataset
- Explainability:
  - Original model
  - Global and per-group
  - Forecast horizon
  - Feature importance
    - Per-group
    - Local Shapley values

# MLI Cheatsheet




https://github.com/h2oai/mli-resources/blob/master/cheatsheet.png

# MLI Functional Architecture

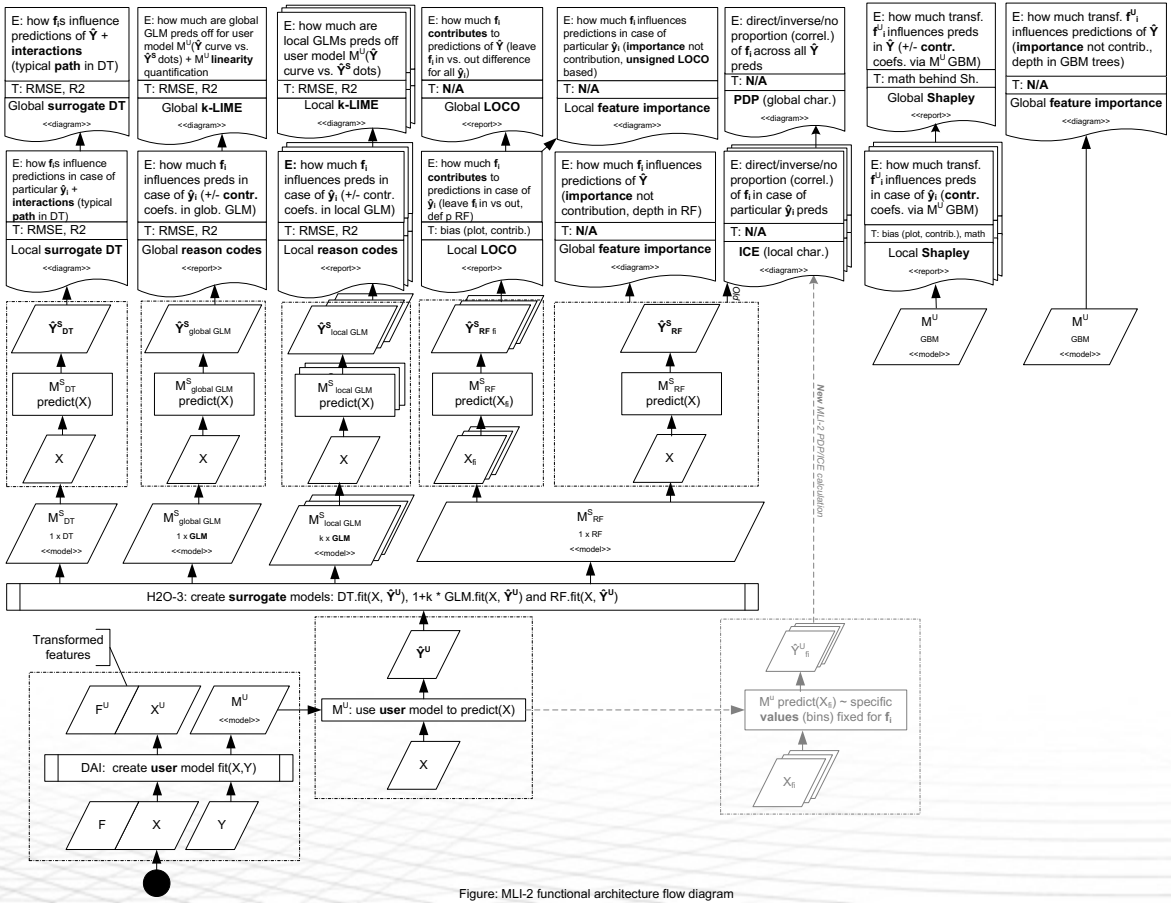



Figure: MLI-2 functional architecture flow diagram

# Conclusion

*Takeaways*

# TAKEAWAYS

- ML interpretability **matters**.
- **Multiplicity** of good models.
- H2O Driverless AI has **interpretability**.
- **Control** model interpretability **end to end**.
- Prefer **interpretable models**.
- **Test** both your model and explanatory SW.
- Use synergy of **local** & **global** techniques.
- **Shapley** values.

# MLI TEAM

Patrick

Navdeep

Mateusz

Zac

Laco

Martin

# Thank you!

# Resources

*Books, articles, links and Git repos,*

# Resources

- https://www.h2oai.com/explainable-ai/
- Booklets:
  - Machine Learning Interpretability with DAI
  - Ideas on Interpreting Machine Learning
- Driverless AI's MLI module cheatsheet
- MLI presentations:
  - MLI walkthrough by Patrick Hall
  - Human Friendly Machine Learning by Patrick Hall
- GitHub repositories:
  - MLI Resources
  - H2O Meetups