

Automatic Feature Engineering in H₂O Driverless AI



Jo-fai (Joe) Chow
Data Science Evangelist /
Community Manager

joe@h2o.ai
@matlabulous

More Info → [https://bit.ly/
h2o_meetups](https://bit.ly/h2o_meetups)

About Me



Jo-fai (Joe) Chow

Data Science Evangelist & Community Manager joe@h2o.ai

- Before H₂O
 - Water Engineer / EngD Researcher / Matlab Fan Boy
(wonder why  @matlabulous?)
 - Discovered R, Python, H₂O ...
never look back again
 - Data Scientist at Virgin Media (UK),
Domino Data Lab (US)
 - At H₂O ...
 - Data Scientist / Evangelist /
 - Sales Engineer / Solution Architect /
 - Event Organiser
 - Photographer

... The harsh reality of startup life ...

Driverless AI Delivers “Expert Data Scientist in a Box”

- Created and supported by world renowned AI experts
- Empowers companies to accomplish AI and ML with a single platform
- Performs the function of an expert data scientist and adds more power to both novice and expert teams
- Details and highlights insights and interpretability with easy to understand results and visualizations



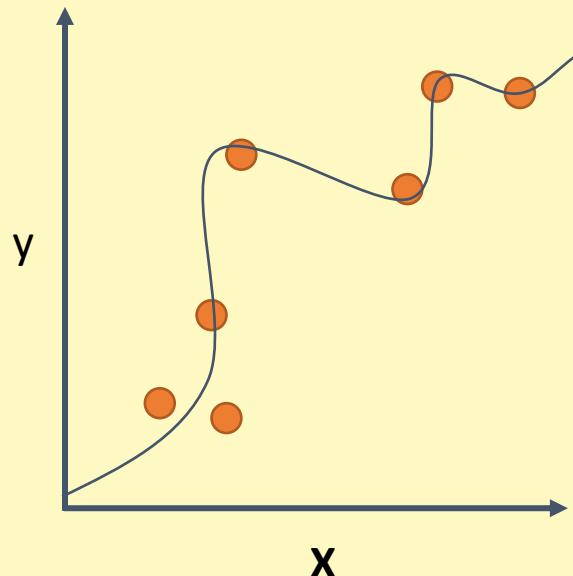
21 day free trial for [Driverless AI](#)

H₂O.ai

Supervised Learning

Regression:

How much will a customer spend?

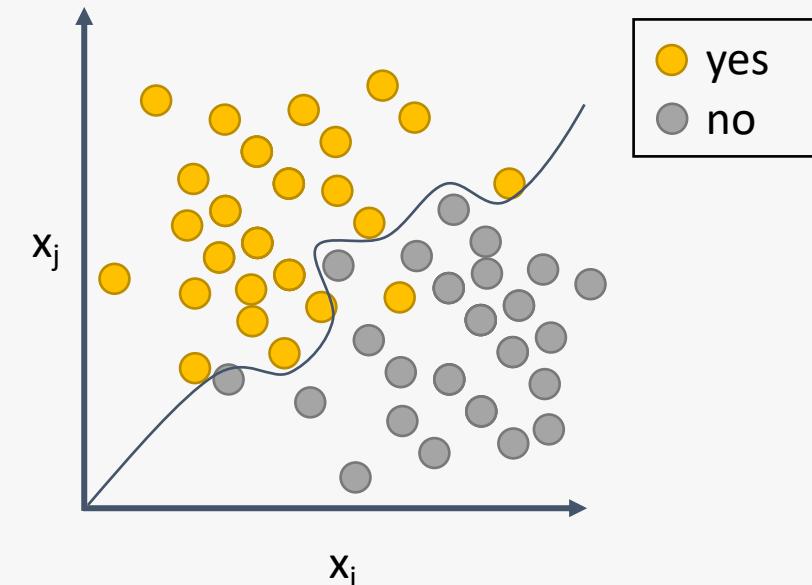


H₂O algos:

Penalized Linear Models
Random Forest
Gradient Boosting
Neural Networks
Stacked Ensembles

Classification:

Will a customer churn?

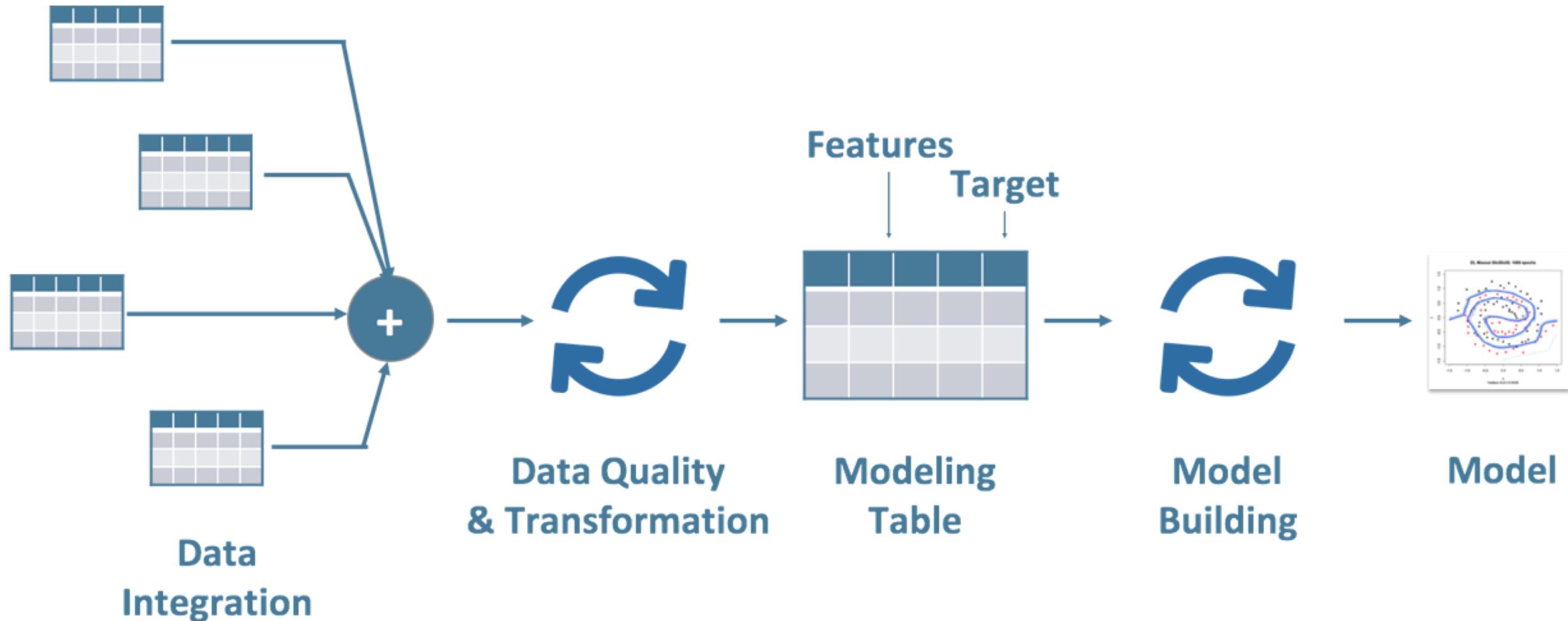


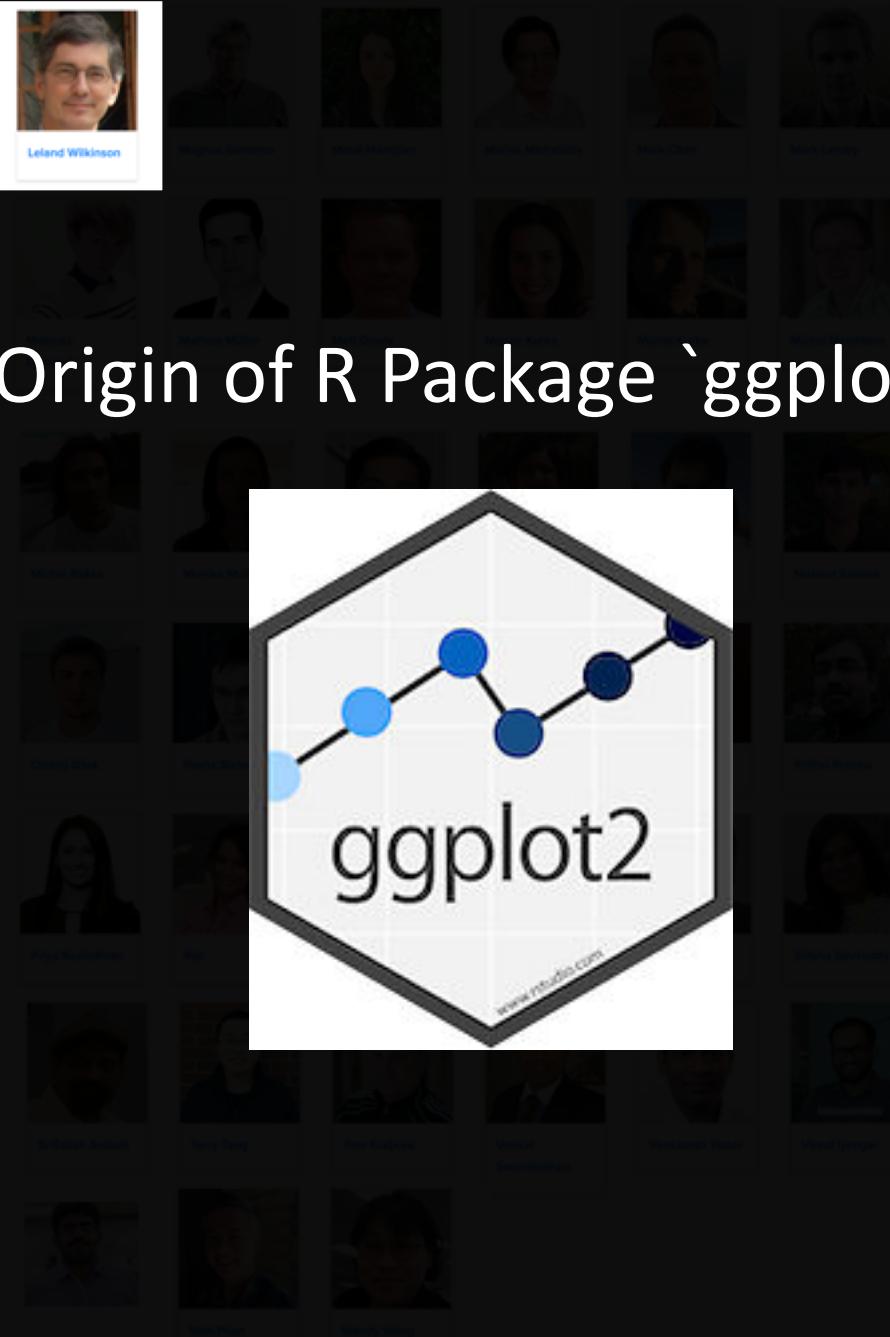
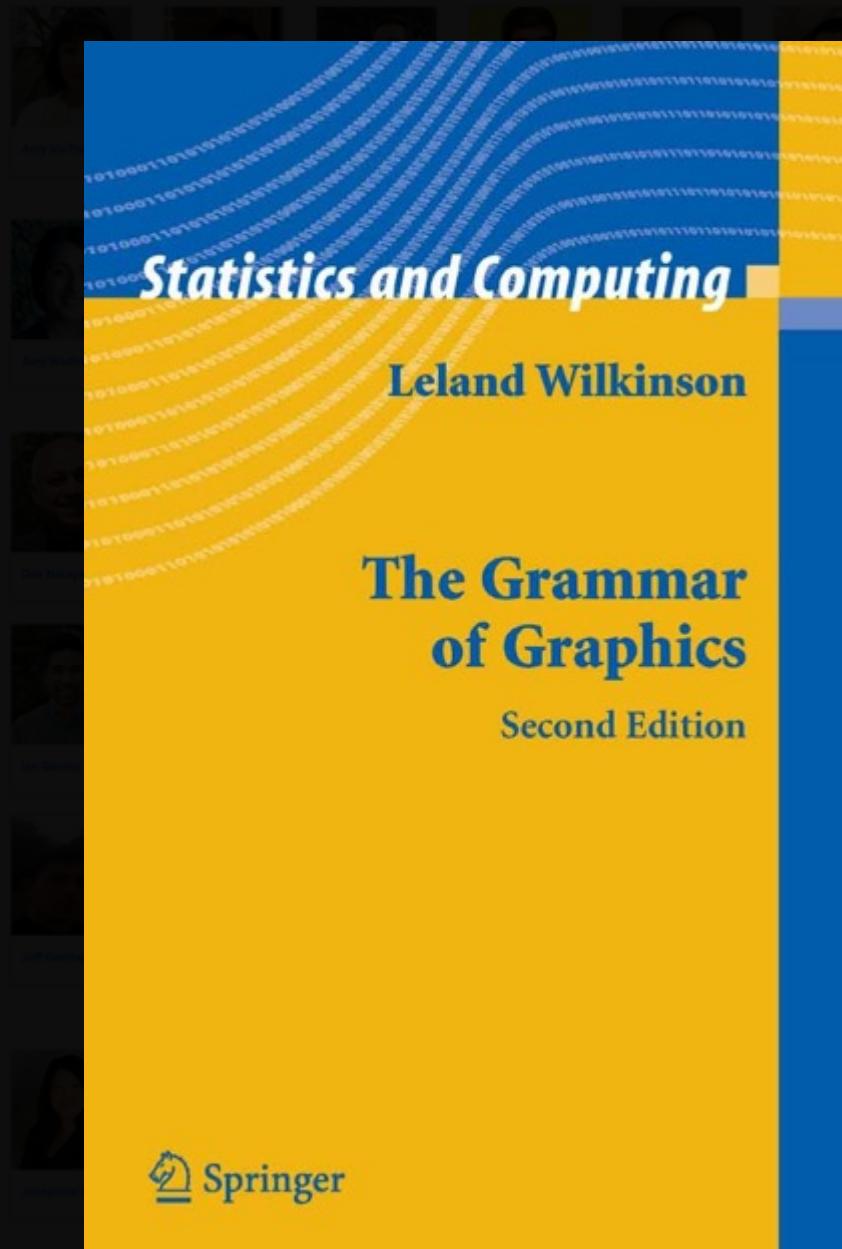
H₂O algos:

Penalized Linear Models
Naïve Bayes
Random Forest
Gradient Boosting
Neural Networks
Stacked Ensembles

H₂O.ai

Driverless AI: Automates Data Science and ML Workflows



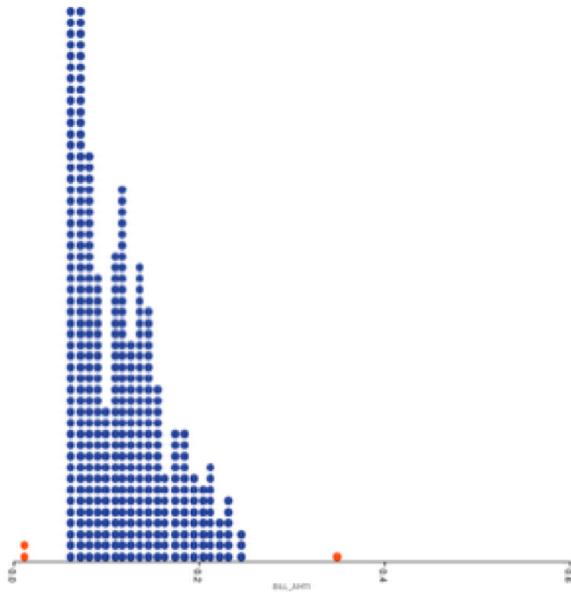


Origin of R Package `ggplot2`

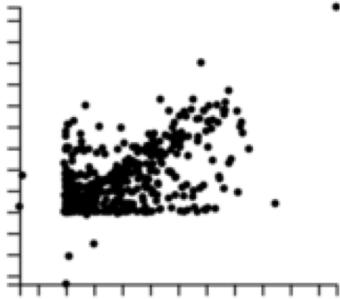
Automatic Visualization

H2O.ai

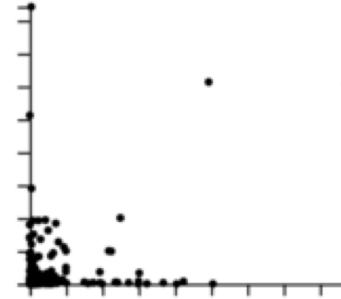
Automatic Scagnostics and other visualizations to generate the most relevant visualizations for each dataset



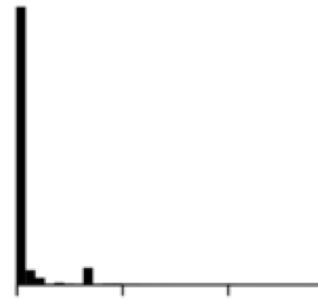
CLUMPY SCATTERPLOTS



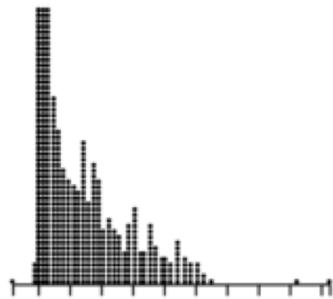
MONOTONIC SCATTERPLOTS



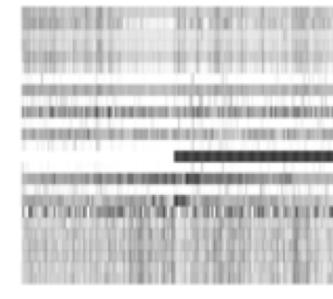
SPIKEY HISTOGRAMS

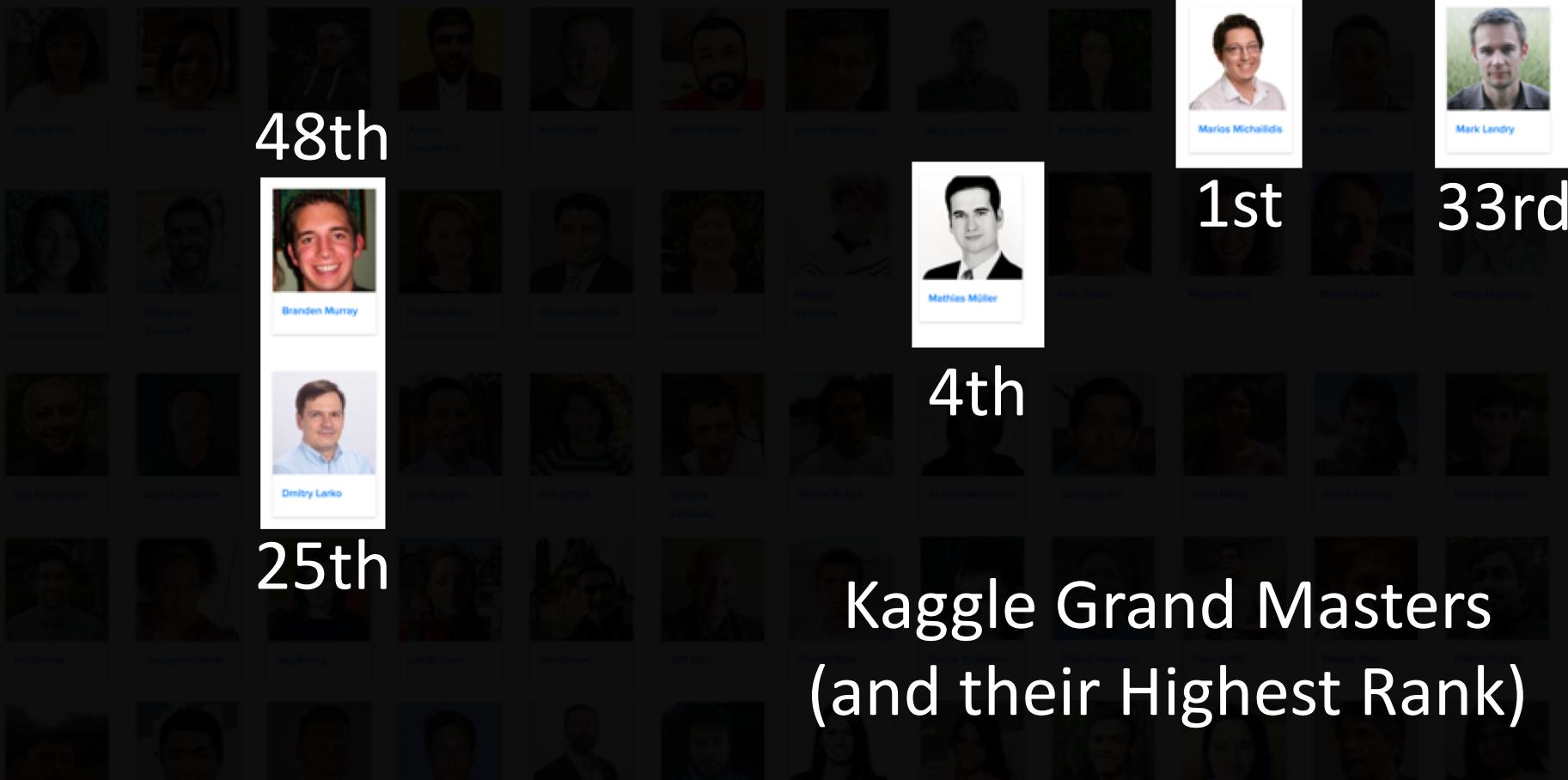


OUTLIERS



HEATMAPS





Kaggle Grand Masters (and their Highest Rank)

 **113**
Grandmasters

 **980**
Masters

 **3,339**
Experts

 **46,135**
Contributors

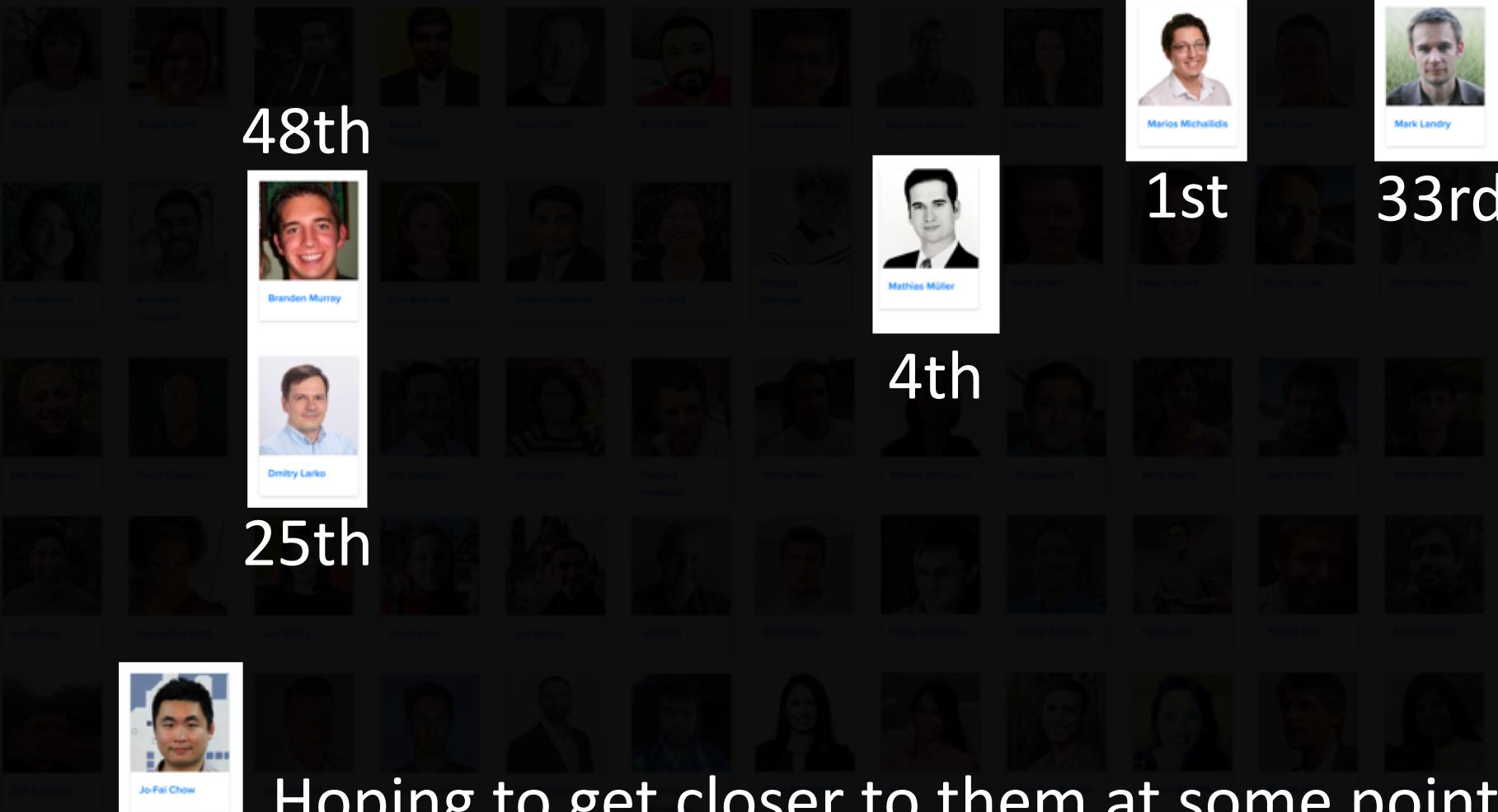
 **33,242**
Novices

About 80,000 Kagglers

H₂O Team

13th

H₂O.ai



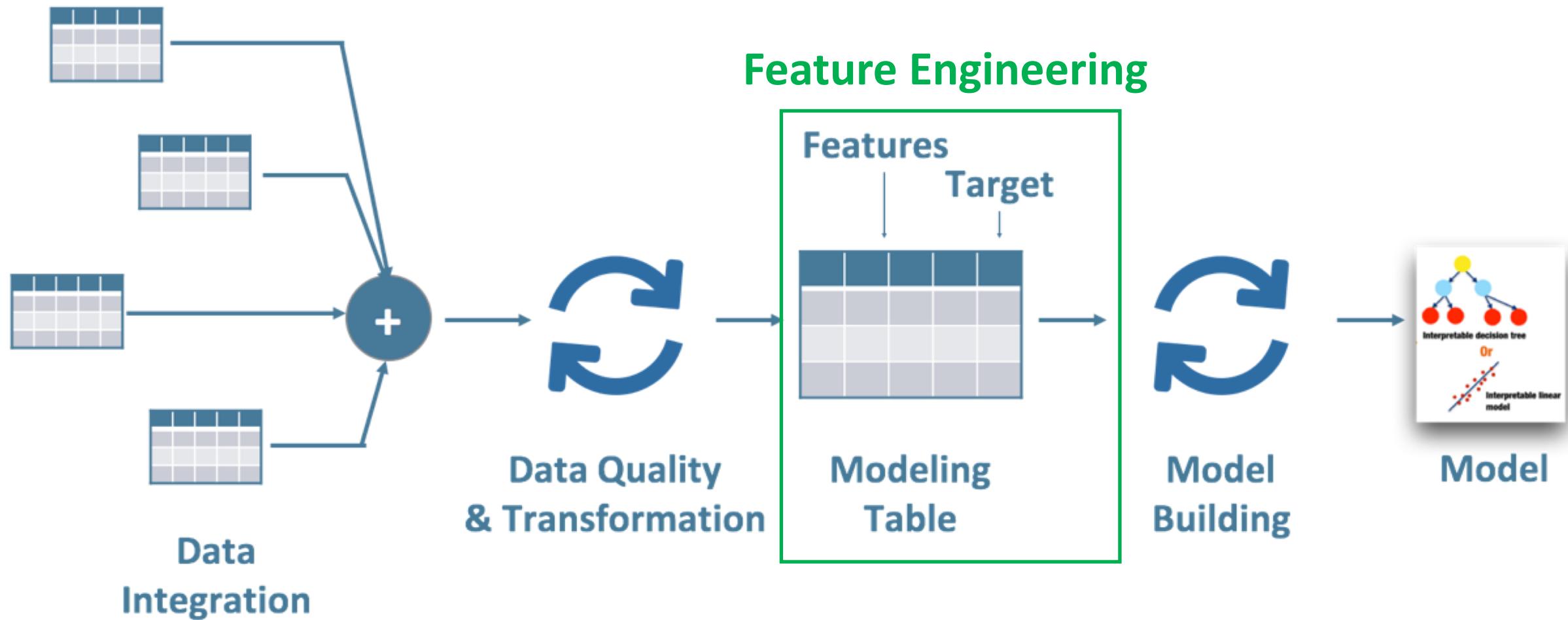
Hoping to get closer to them at some point ...

H₂O Team

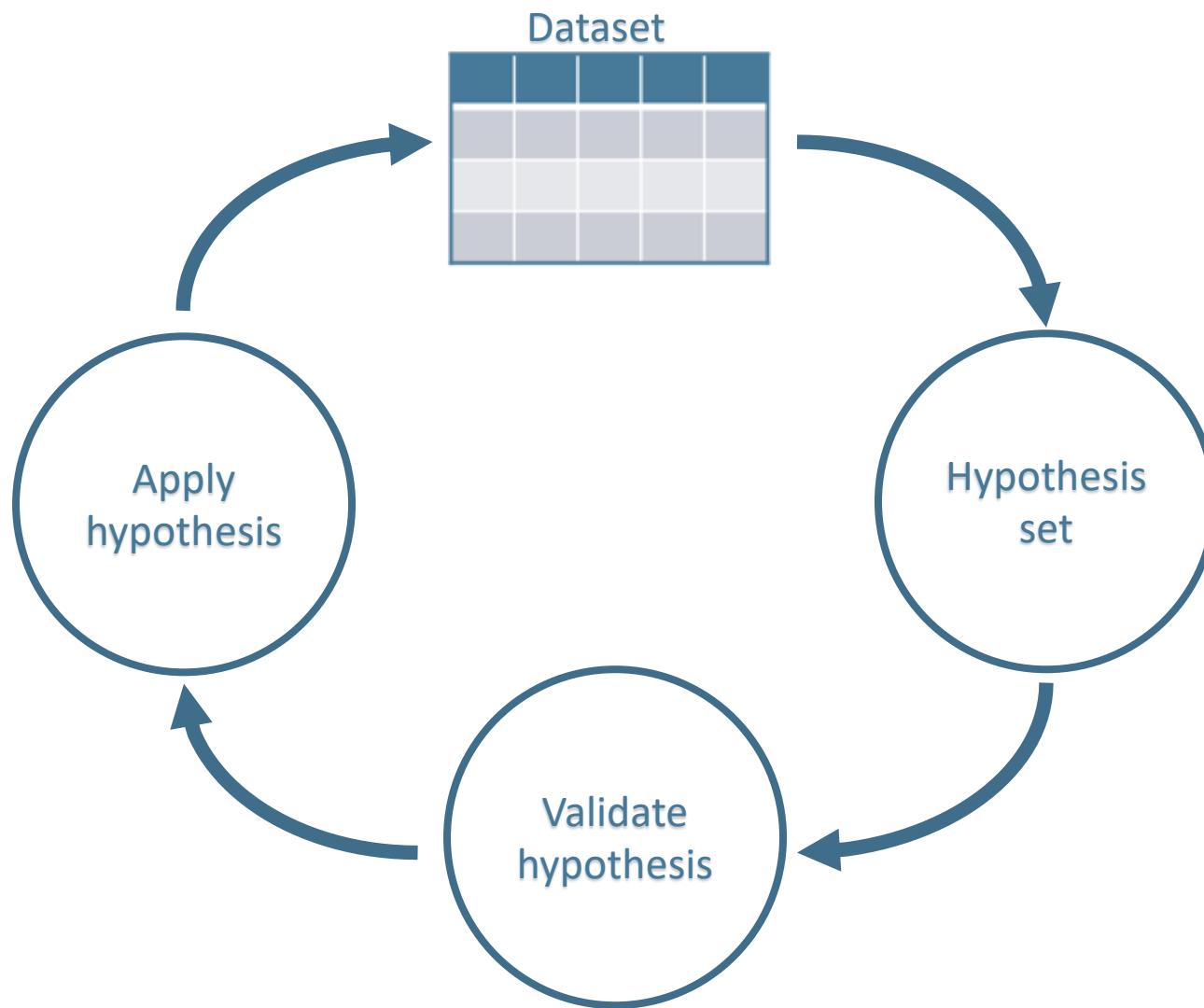
13th

H₂O.ai

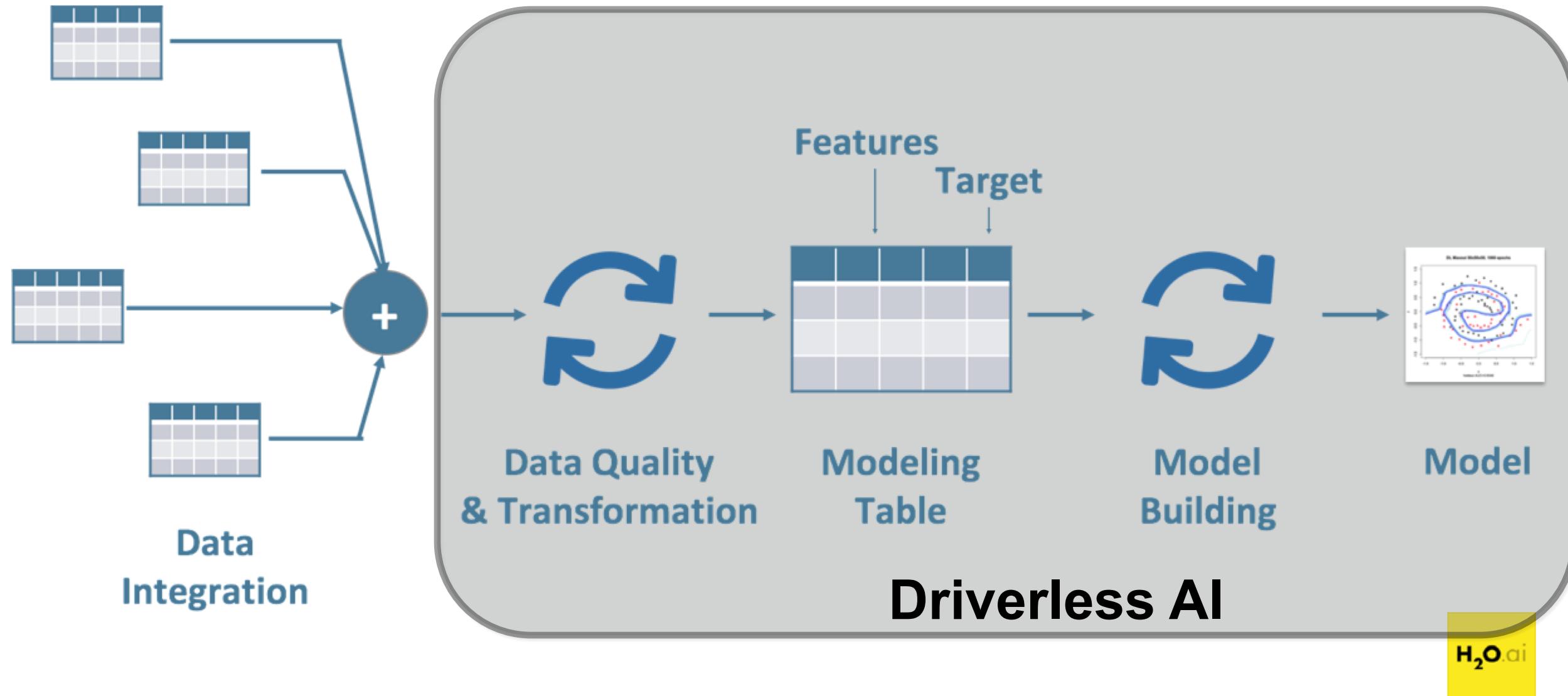
Typical Enterprise Machine Learning Workflow



Feature Engineering cycle



Driverless AI: Automates Data Science and ML Workflows



Accuracy

- Automatic feature engineering to increase accuracy - AlphaGo for AI
- Automatic Kaggle Grandmaster recipes in a box for solving wide variety of use-cases
- Automatic machine learning to find and tune the right ensemble of models

Driverless AI: top 5% in Amazon Kaggle competition

Driverless AI produces feature engineering pipeline (“more columns”) for downstream use



Amazon.com - Employee Access

Predict an employee's access needs
\$5,000 - 1,687 teams - 4 years ago

Driverless AI: 80th place (out of 1687 - top 5%)

Driverless AI: Top-10 in BNP Paribas Kaggle competition



single run, fully automated: 2h on DGX Station! 6h on PC

BNP Paribas Cardif Claims Management

Can you accelerate BNP Paribas Cardif's claims management process?
\$30,000 - 2,926 teams - 2 years ago

Submission and Description
[sub.csv](#)
2 minutes ago by Arno Candel
9408bf7 7/10/1 cv 0.4354 finished after 172 iters

Private Score: 0.42945
Public Score: 0.43156

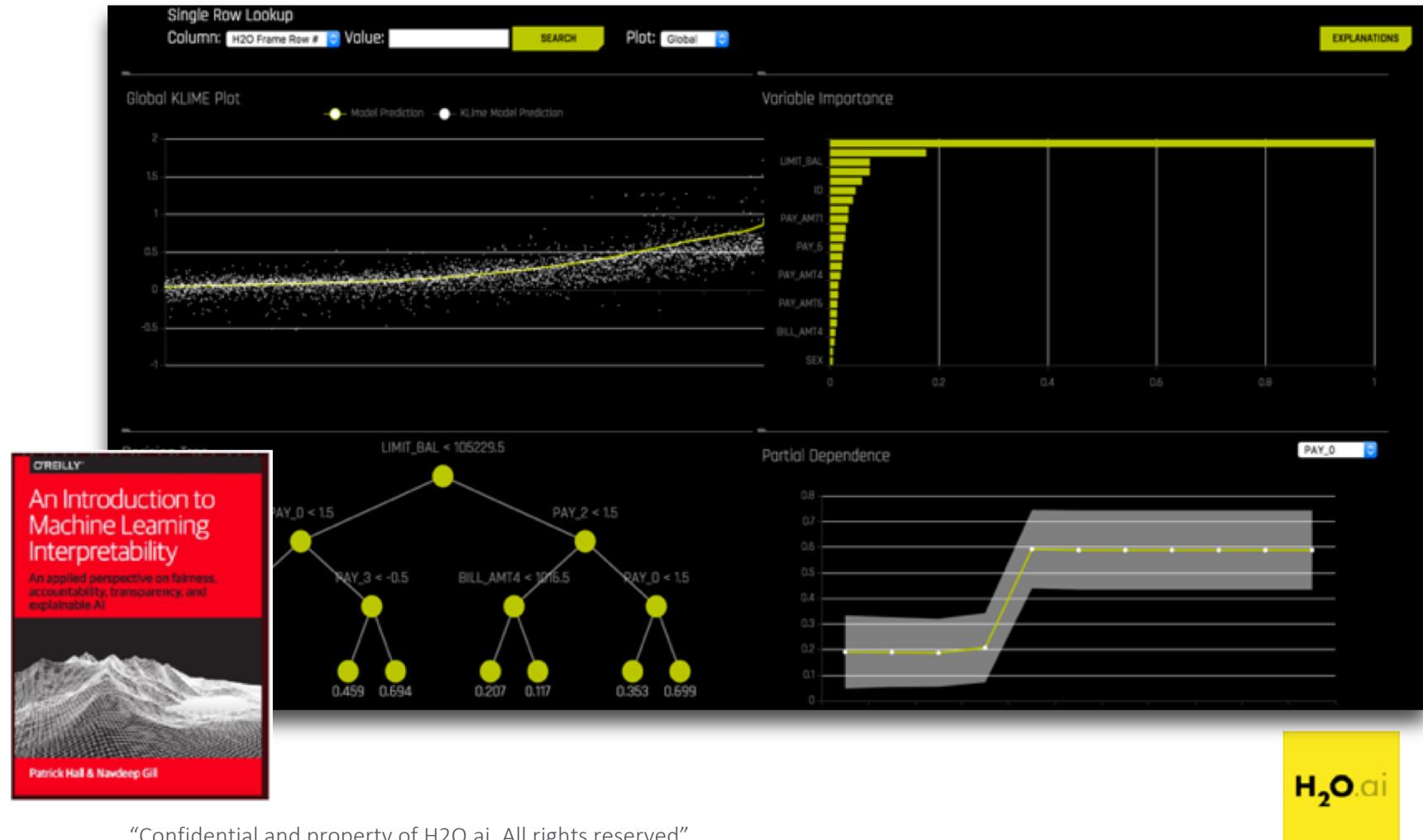
Driverless AI: 10th place in private LB at Kaggle (out of 2926)

2 months for Grandmasters — 2 hours for Driverless AI



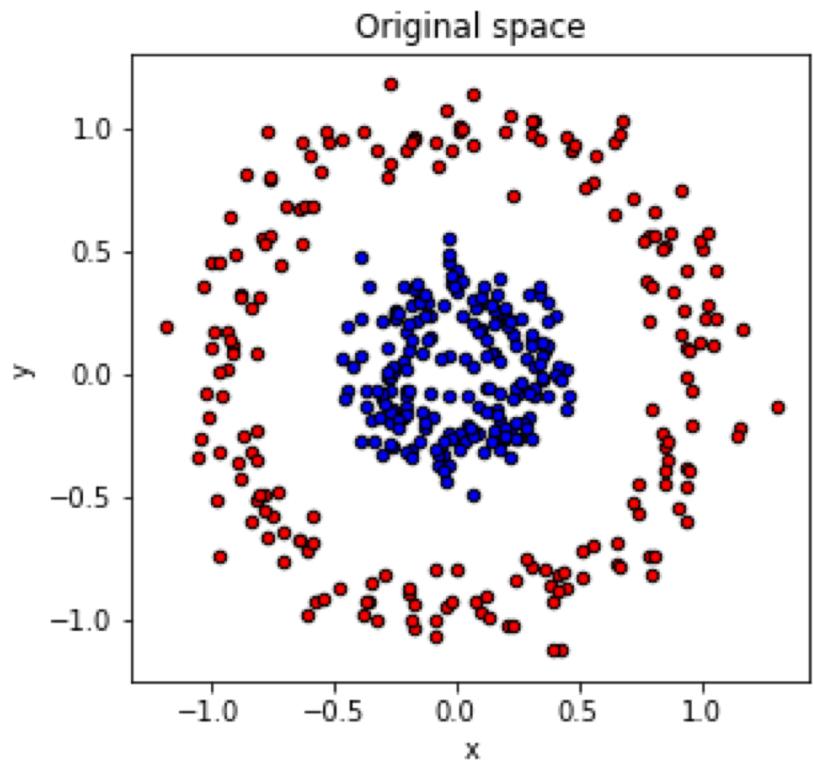
Interpretability

- Interpretability for debugging, not just for regulators
- Get reason codes and model interpretability in plain english
- K-Lime, LOCO, partial dependence and more



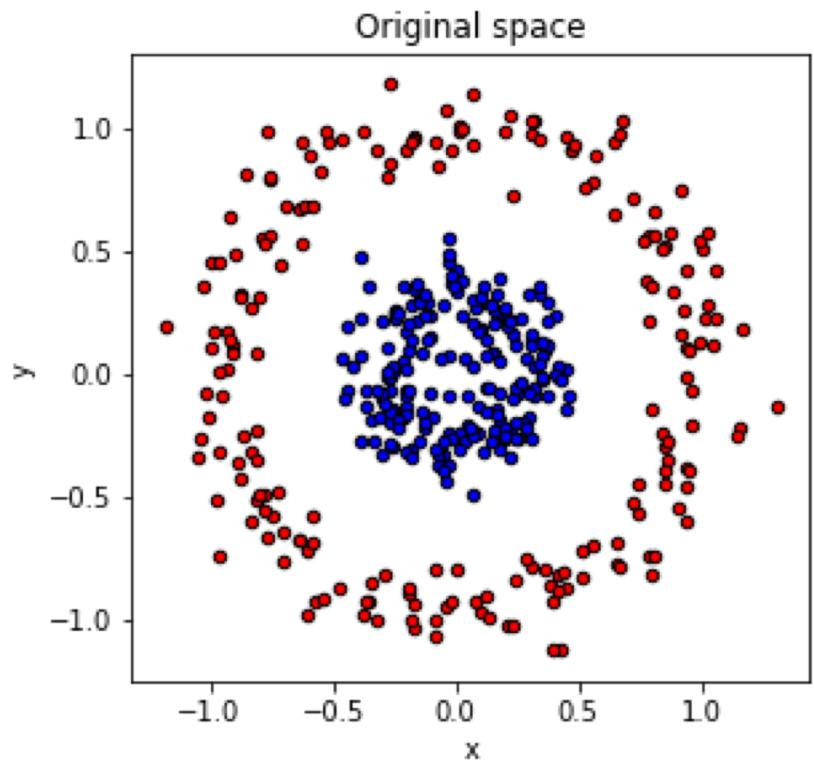
What is Feature Engineering?

What is feature engineering

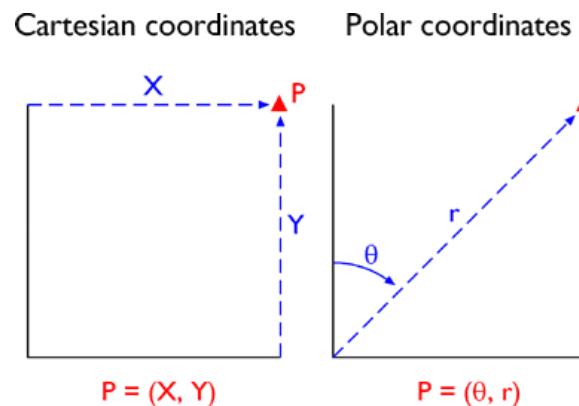


Not possible to separate using linear classifier

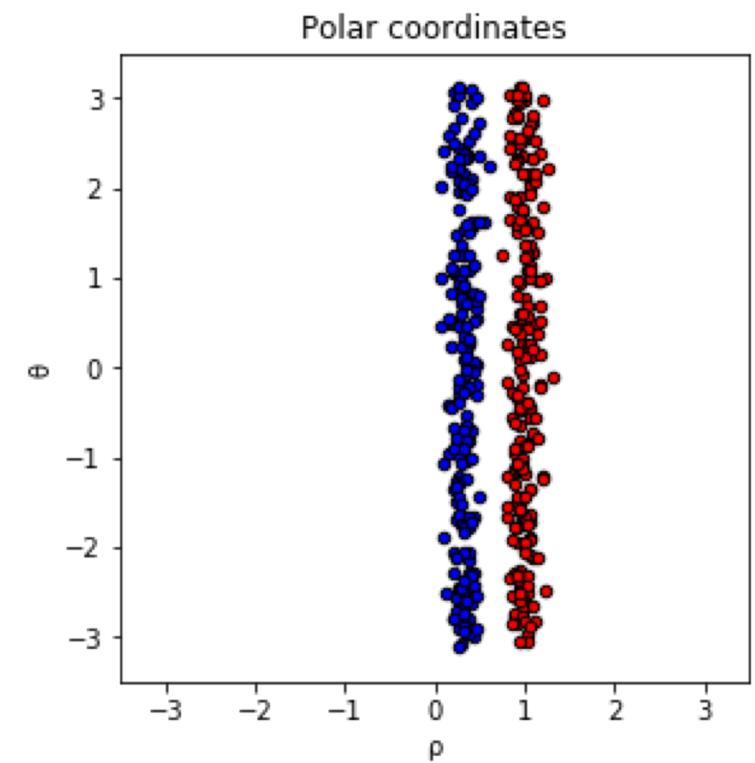
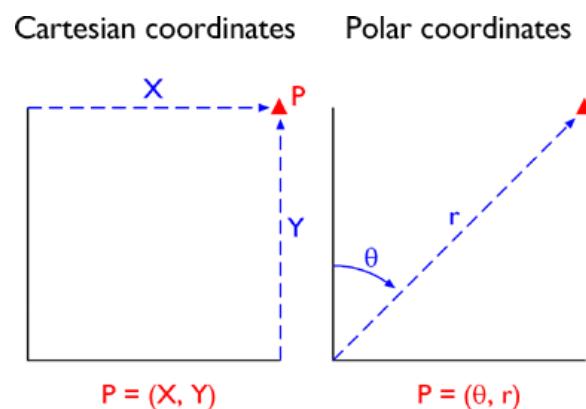
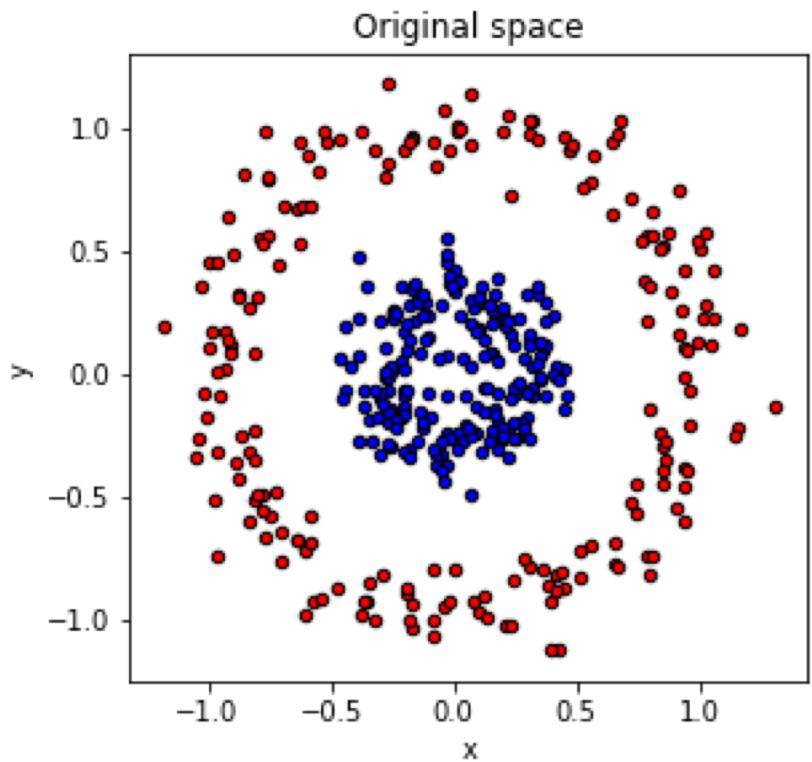
What is feature engineering



What if we use polar coordinates instead?



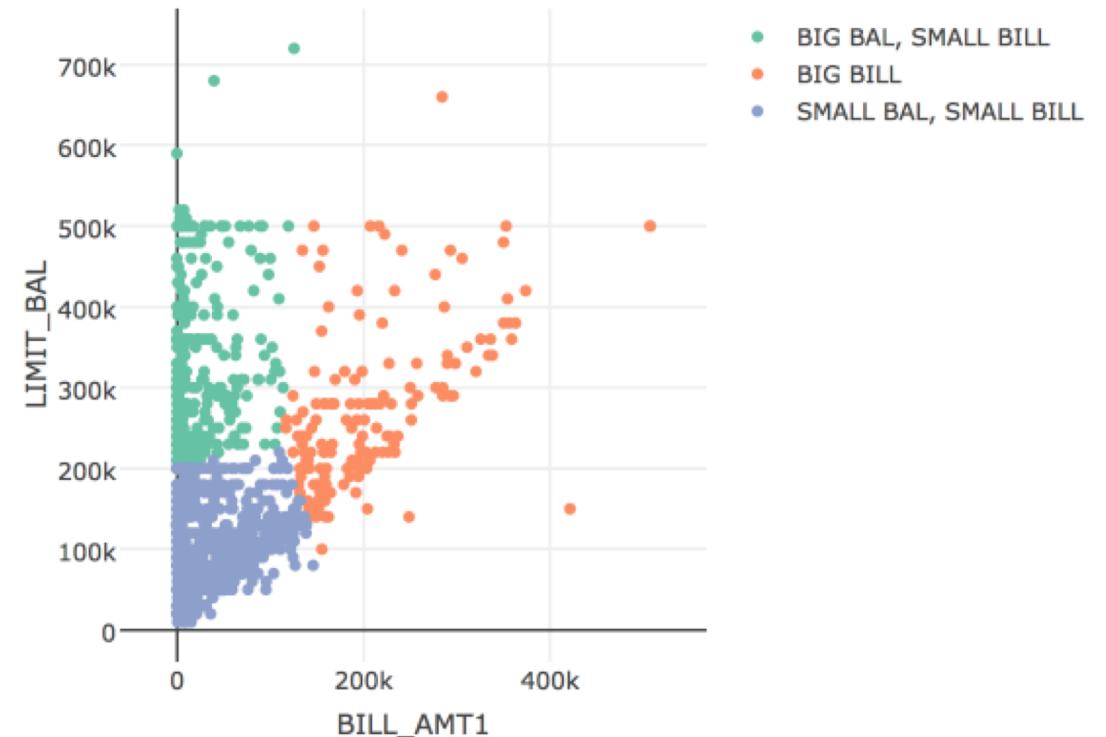
What is feature engineering



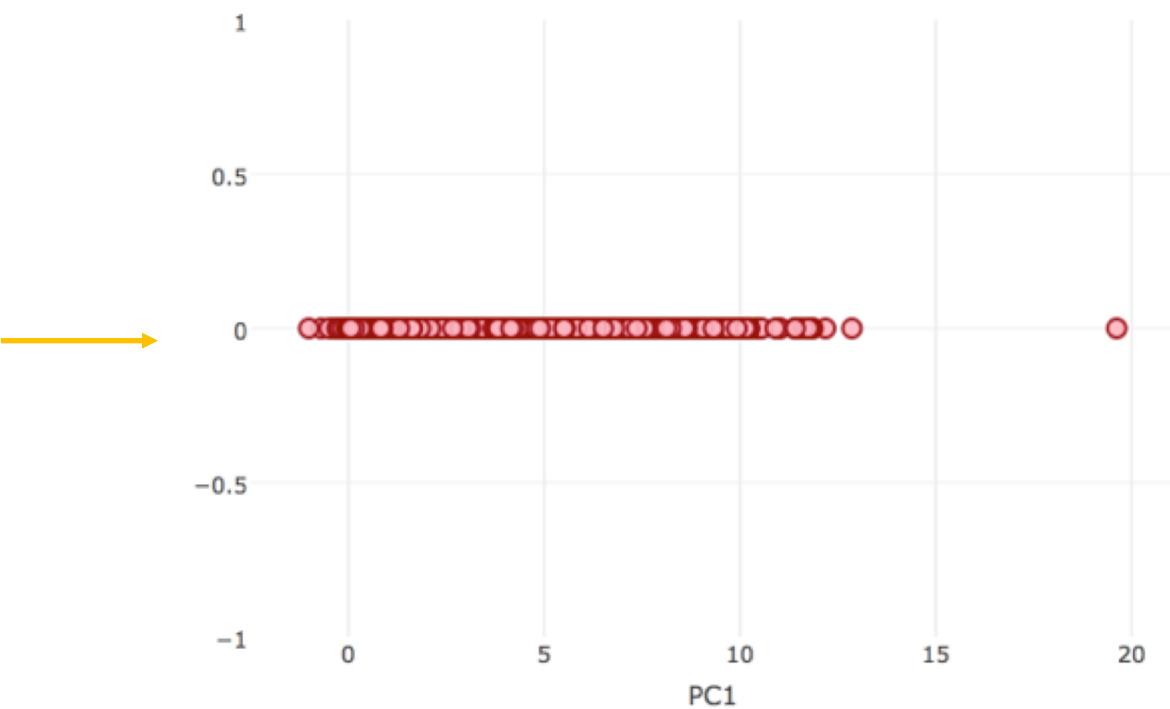
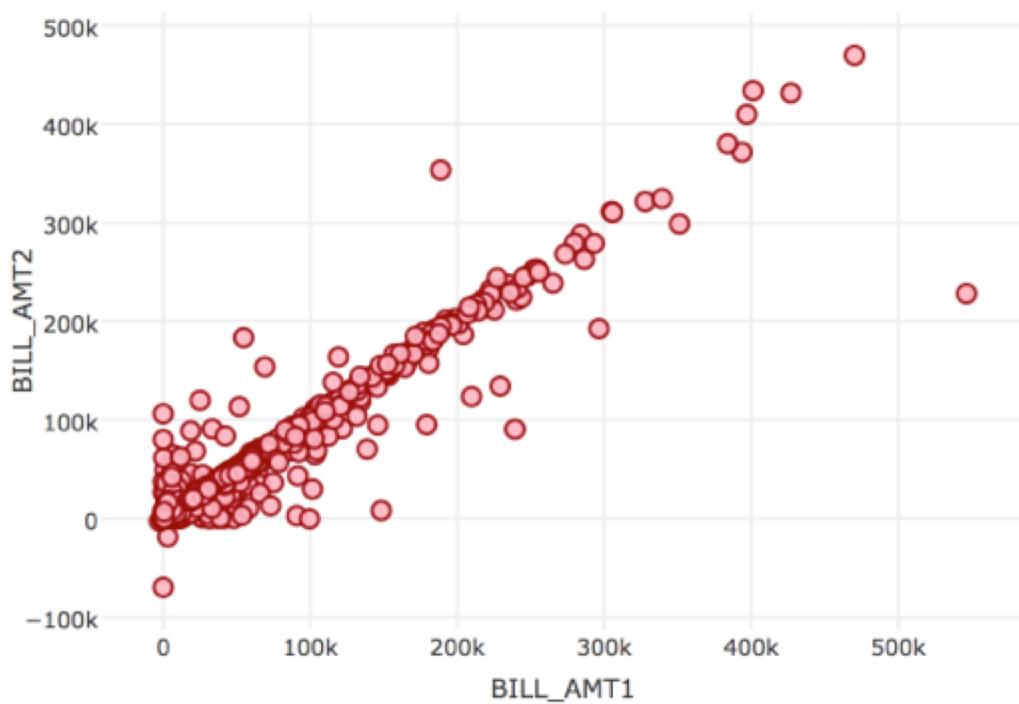
Clustering

Cluster Transformations

- Distance to a specific cluster
- Cross Validation Target Encoding by Cluster ID



Truncated SVD



Text Features

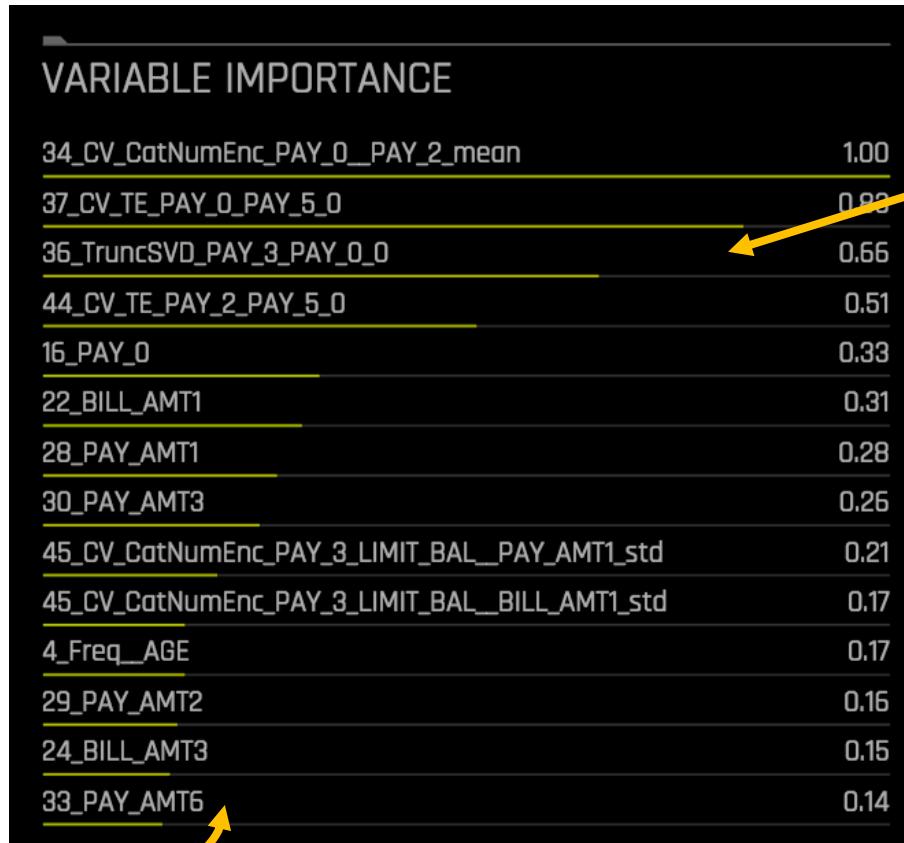
9_TxtTE:Description.0	1.00
27_WoE:HelpfulnessNumerator:Summary.0	0.31
20_WoE:HelpfulnessDenominator:Summary:UserId.0	0.20
4_CVTE:Summary.0	0.18
24_InteractionSub:HelpfulnessDenominator:Helpfu...	0.17
28_ClusterTE:ClusterID70:HelpfulnessDenominator:...	0.15
10_Txt:Description.22	0.05
10_Txt:Description.3	0.05
2_CVTE:ProductId.0	0.04
10_Txt:Description.5	0.03
10_Txt:Description.8	0.03
6_HelpfulnessDenominator	0.03
10_Txt:Description.18	0.03
10_Txt:Description.11	0.03

TxtTE – Train a linear model on the text components from TF-IDF

Txt – Components from a TF-IDF Matrix

There are Many More Tricks!

Kaggle Grand Master Out of the Box



Generated Features

Feature Transformations

- Automatic Text Handling
- Frequency Encoding
- Cross Validation Target Encoding
- Truncated SVD
- Clustering and more

How about Auto Feature Engineering + Marios' StackNet?

Driverless AI + StackNet (9 months ago)

The screenshot shows the Kaggle homepage with the navigation bar: kaggle, Search kaggle, Competitions, Datasets, Kernels, Discussion, Jobs, and a user profile icon. A banner at the top highlights the "Featured Prediction Competition: Zillow Prize: Zillow's Home Value Prediction (Zestimate)." The banner text reads: "Can you improve the algorithm that changed the world of real estate?" It shows "Zillow · 3,779 teams · 2 days ago" and a "\$1,200,000 Prize Money".

Competition Round One (Top 100 to Next Round)

40	▼ 8	Deal or No Deal		0.0749020	79	3mo
41	▲ 52	SCC		0.0749052	39	3mo
42	▼ 31	KFP		0.0749066	349	3mo

Finished above my H2O Kaggle Grandmasters colleagues

The screenshot shows the Kaggle Leaderboard for the "Zillow Prize: Zillow's Home Value Prediction (Zestimate)" competition. The top banner indicates "2 days · 3,779 teams · 2 days ago". The page includes tabs for Overview, Date, Kernels, Discussion, Leaderboard, Rules, Team, My Submissions, and Late Submission. The Leaderboard section shows the public leaderboard with 100 entries. The columns include Rank, Team Name, Kernel, Team Members, Score ID, Entries, and Last. The top entry is "combine_0374sc7v" with a score of 0.0749081. The page also includes a note about the private leaderboard and a refresh button.

Live Demo


BNP PARIBAS CARDIF

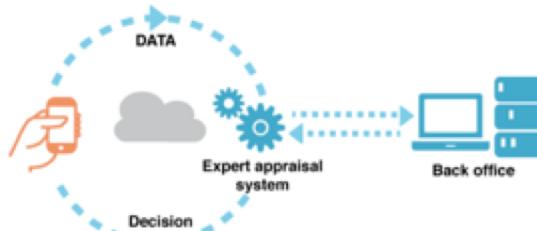
BNP Paribas Cardif Claims Management

Can you accelerate BNP Paribas Cardif's claims management process?
\$30,000 · 2,926 teams · 2 years ago

Overview Data Kernels Discussion Leaderboard Rules Team My Submissions Late Submission

Overview

Description	As a global specialist in personal insurance, BNP Paribas Cardif serves 90 million clients in 36 countries across Europe, Asia and Latin America.
Evaluation	In a world shaped by the emergence of new uses and lifestyles, everything is going faster and faster. When facing unexpected events, customers expect their insurer to support them as soon as possible. However, claims management may require different levels of check before a claim can be approved and a payment can be made. With the new practices and behaviors generated by the digital economy, this process needs adaptation thanks to data science to meet the new needs and expectations of customers.
Prizes	
Timeline	
About Bnp Paribas Cardif	



In this challenge, BNP Paribas Cardif is providing an anonymized database with two categories of claims:

1. claims for which approval could be accelerated leading to faster payments
2. claims for which additional information is required before approval

Kagglers are challenged to predict the category of a claim based on features available early in the process, helping BNP Paribas Cardif accelerate its claims process and therefore provide a better service to its customers.

LINK: <https://techdayhq.com/london/register#attend>

REGISTRATION CODE: **LDNFREE**





Turner & Townsend

H₂O.ai

- More Info, Code, and Slides
 - [bit.ly/
h2o_meetups](https://bit.ly/h2o_meetups)
- Contact
 - joe@h2o.ai
 - [@matlabulous](https://twitter.com/matlabulous)
 - github.com/woobe

Appendix

Cross Validation Target Encoding

Fold	Pay 1	Default Payment
1	Up To Date	0
2	Up To Date	0
3	Up To Date	0
2	Missed 1 Mo	1
1	Missed 1 Mo	0
3	Missed 1 Mo	0
1	Missed 5 Mo	1

Cross Validation Target Encoding

Fold	Pay 1	Default Payment
2	Up To Date	0
3	Up To Date	0
2	Missed 1 Mo	1
3	Missed 1 Mo	0

Fold	Pay 1	Default Payment
1	Up To Date	0
1	Missed 1 Mo	0
1	Missed 5 Mo	1

Cross Validation Target Encoding

Fold	Pay 1	Default Payment	Mean Target Encoding
2	Up To Date	0	0
3	Up To Date	0	0
2	Missed 1 Mo	1	0.5
3	Missed 1 Mo	0	0.5

Fold	Pay 1	Default Payment
1	Up To Date	0
1	Missed 1 Mo	0
1	Missed 5 Mo	1

Cross Validation Target Encoding

Fold	Pay 1	Default Payment	Mean Target Encoding
2	Up To Date	0	0
3	Up To Date	0	
2	Missed 1 Mo	1	0.5
3	Missed 1 Mo	0	

Fold	Pay 1	Default Payment	CV Target Encoding
1	Up To Date	0	0
1	Missed 1 Mo	0	
1	Missed 5 Mo	1	

Cross Validation Target Encoding

Fold	Pay 1	Default Payment	Mean Target Encoding
2	Up To Date	0	0
3	Up To Date	0	0
2	Missed 1 Mo	1	0.5
3	Missed 1 Mo	0	0.5

Fold	Pay 1	Default Payment	CV Target Encoding
1	Up To Date	0	0
1	Missed 1 Mo	0	0.5
1	Missed 5 Mo	1	

Cross Validation Target Encoding

Fold	Pay 1	Default Payment	Mean Target Encoding
2	Up To Date	0	0
3	Up To Date	0	0
2	Missed 1 Mo	1	0.5
3	Missed 1 Mo	0	0.5

Fold	Pay 1	Default Payment	CV Target Encoding
1	Up To Date	0	0
1	Missed 1 Mo	0	0.5
1	Missed 5 Mo	1	NA