Introduction to H_2O Driverless AI

Use Cases and Live Demos

FinTech Scotland 2018

In association with VisitScotland Business Events

Jo-fai (Joe) Chow

Data Science Evangelist / Community Manager

joe@h2o.ai

@matlabulous

More Info → https://bit.ly/

h2o_meetups

About Me

LONDON R
Tuesday 11th March 2014

10:53 PM - 30 Nov 2017 from Glasgow, Scotland

Glasgow 2017

11:34 PM - 15 Aug 2018 from Glasgow, Scotland

Glasgow 2018

- **Before H_2O**
  - Water Engineer / EngD Researcher / Matlab Fan Boy
    (wonder why @matlabulous?)
  - Discovered R, Python, H_2O … never look back again
  - Data Scientist at Virgin Media (UK), Domino Data Lab (US)
- **At H_2O …**
  - Data Scientist / Evangelist /
  - Sales Engineer / Solution Architect /
  - Community Manager

  … The harsh reality of startup life …

Reminder: #360Selfie

H_2O.ai

H2O.ai Overview

Company	Founded in Silicon Valley in 2012 Funded: $75M Investors: Wells Fargo, NVIDIA, Nexus Ventures, Paxion Ventures
Products	• H2O Open Source Machine Learning (14,000 organizations) • H2O Driverless AI – Automatic Machine Learning
Leadership	Leader in Gartner MQ Machine Learning and Data Science Platform
Team	120 AI experts (Kaggle Grandmasters, Distributed Computing, Visualization)
Global	Mountain View, London, Prague, India

Scientific Advisory Council

Dr. Trevor Hastie

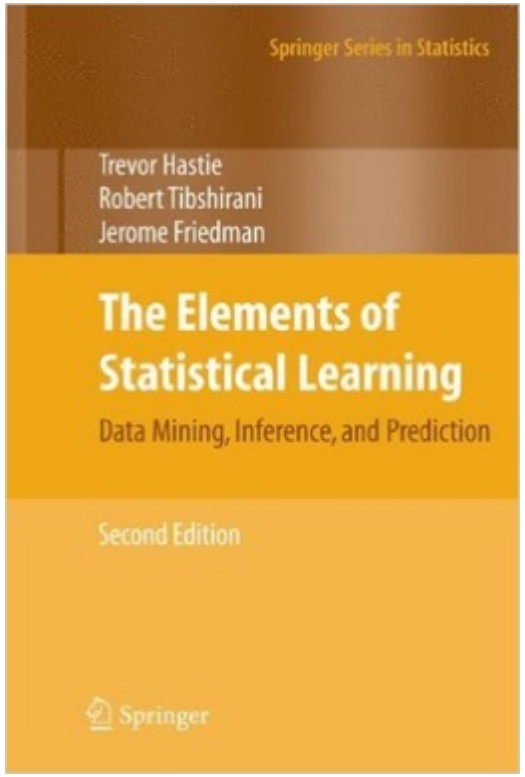
- John A. Overdeck Professor of Mathematics, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Co-author with John Chambers, *Statistical Models in S*
- Co-author, *Generalized Additive Models*

Dr. Robert Tibshirani

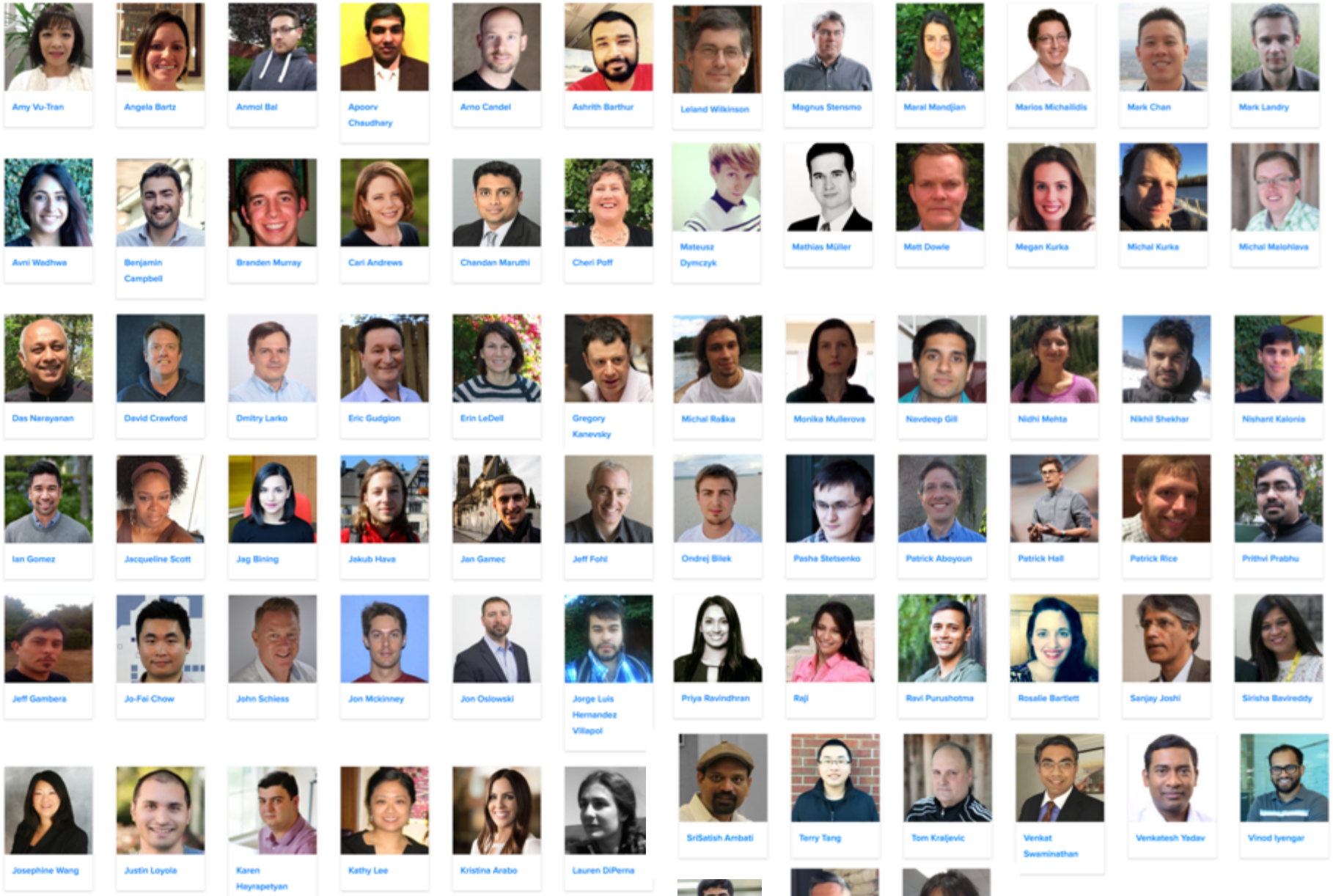
- Professor of Statistics and Health Research and Policy, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Author, *Regression Shrinkage and Selection via the Lasso*
- Co-author, *An Introduction to the Bootstrap*

Dr. Steven Boyd

- Professor of Electrical Engineering and Computer Science, Stanford University
- PhD in Electrical Engineering and Computer Science, UC Berkeley
- Co-author, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*
- Co-author, *Linear Matrix Inequalities in System and Control Theory*
- Co-author, *Convex Optimization*

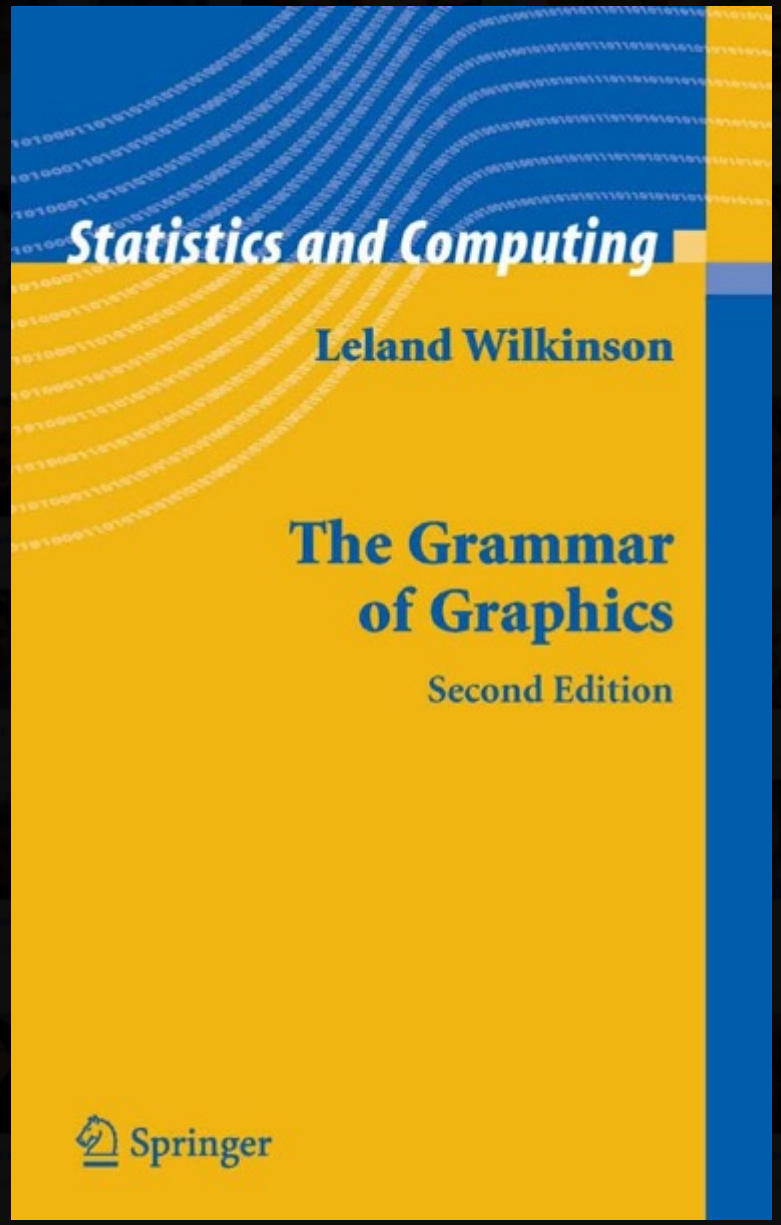


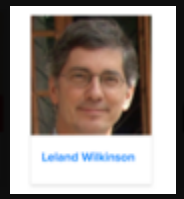
H2O.ai



H_2O Team

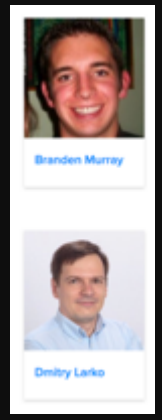
H_2O.ai



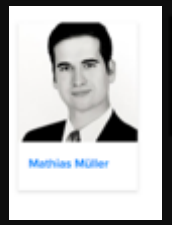


Origin of R Package `ggplot2`

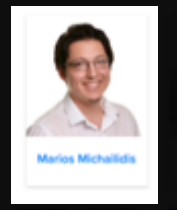
48th



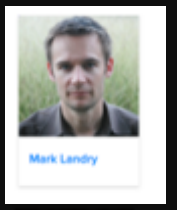
25th



4th



1st



33rd

Kaggle Grandmasters
(and their Highest Rank)

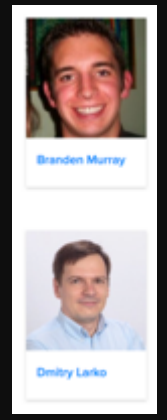
About 80,000 Kagglers

13th

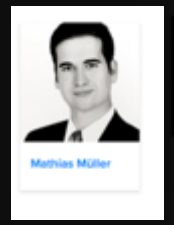
H2O Team

H2O.ai

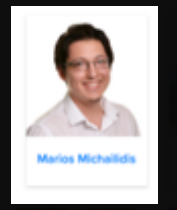
48th



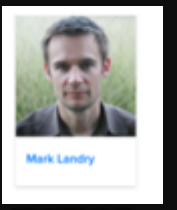
25th



4th



1st



33rd

181st

Hoping to get closer to them at some point …

13th

H2O.ai Product Suite

In-Memory, Distributed Machine Learning Algorithms with H2O Flow GUI

H2O AI Open Source Engine Integration with Spark

Lightning Fast machine learning on GPUs

- 100% open source – Apache V2 licensed
- Built for data scientists – interface using R, Python on H2O Flow (interactive notebook interface)
- Enterprise Support subscriptions

Automatic feature engineering, machine learning and interpretability

- Enterprise software
- Built for domain users, analysts & data scientists – GUI based interface for end-to-end data science
- Fully automated machine learning from ingest to deployment
- User licenses on a per seat basis (annual subscription)

Worldwide Recognition in the H2O.ai Community

Open source community

222 OF FORTUNE 500 ♥ H_2O

8 OF TOP 10 BANKS

7 OF TOP 10 INSURANCE COMPANIES

4 OF TOP 10 HEALTHCARE COMPANIES

Paying Customers

"H2O.ai's reference customers gave it the highest overall score for sales relationship and overall service and support" - Gartner MQ 2018

H2O.ai

Partner Ecosystem

H2O.ai is a Leader in the 2018 Gartner Data Science and Machine Learning Platforms Magic Quadrant

- **Technology leader with most completeness of vision**

- Recognized for the mindshare, partner network and status as a **quasi-industry standard** for machine learning and AI

- **H2O.ai customers gave the highest overall score** among all the vendors for sales relationship and account management, customer support (onboarding, troubleshooting, etc.) and overall service and support

Figure 1. Magic Quadrant for Data Science and Machine-Learning Platforms

Source: Gartner (February 2018)

Get the Gartner Magic Quadrant here

Why Driverless AI?

Why Driverless AI for Enterprise AI adoption

Data is a Team Sport

~100

Data science experts in the world

Lack of AI Talent

Months to Insight

Time for a data scientist to build a model

Time to Insights slow

Black box models

Lack of Trust in AI

”US alone faces a shortage of 190,000 people with analytical expertise.”

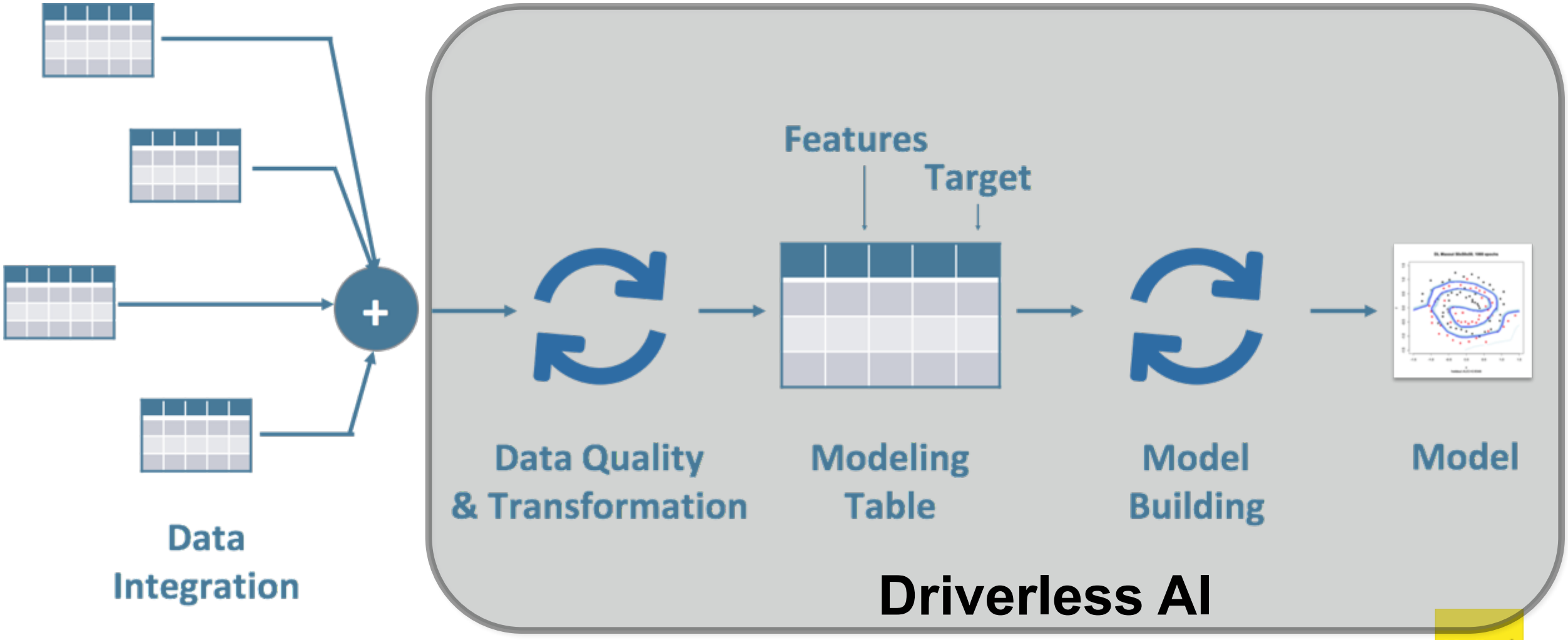
McKinsey&Company

H2O.ai

Driverless AI Delivers AI for Enterprise

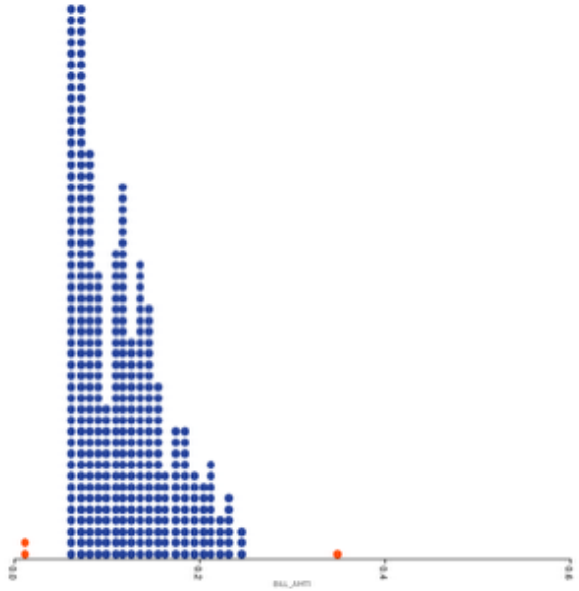
Talent	Time	Trust
Top 10 Data Science Experts	Months to Hours GPU accelerated ML Automatic Pipelines	MLI Auto Doc Auto Visualization
Kaggle Grandmasters	Time to Insight	Explainability & Transparency

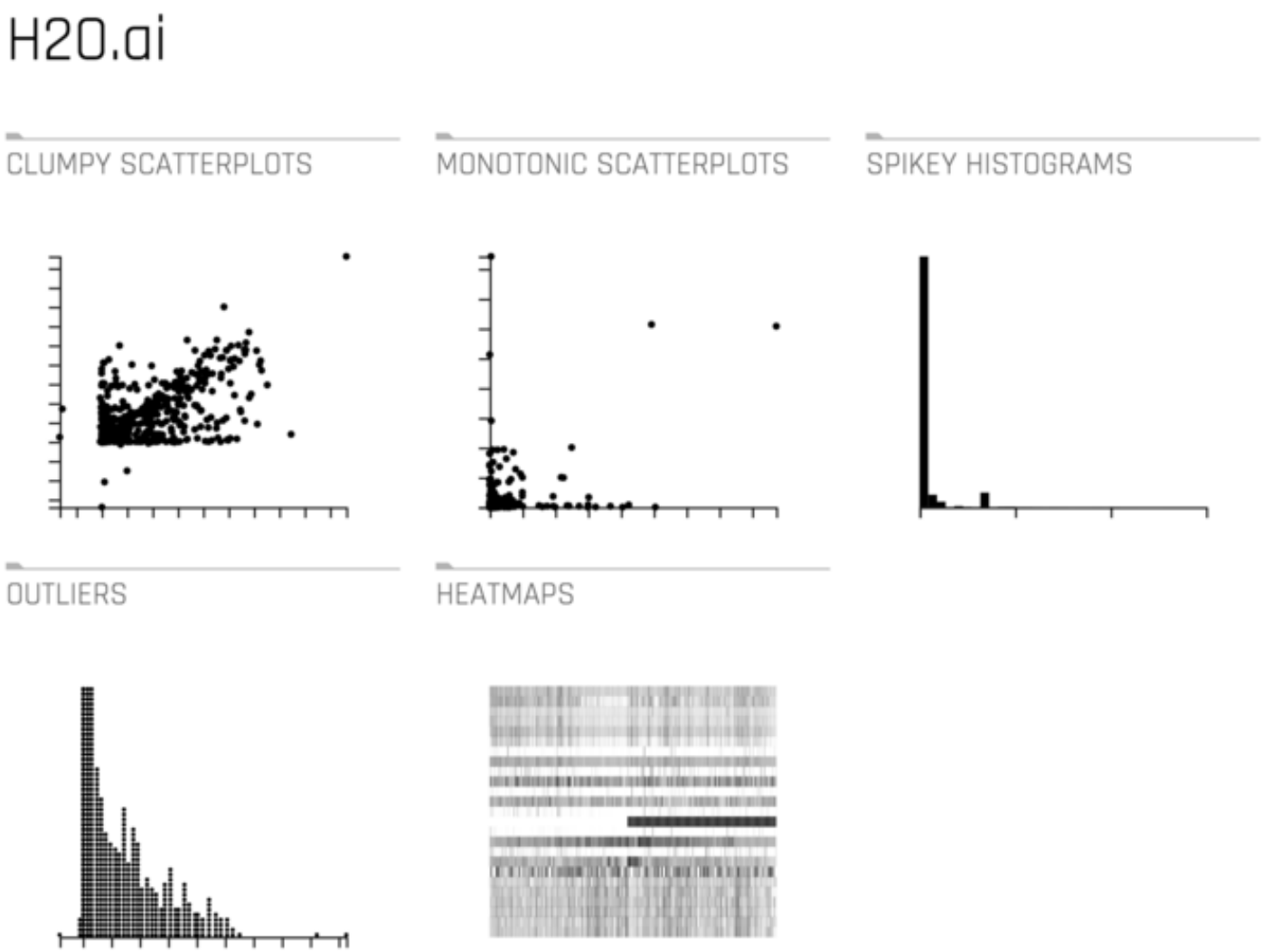
Driverless AI: Automates Data Science and ML Workflows



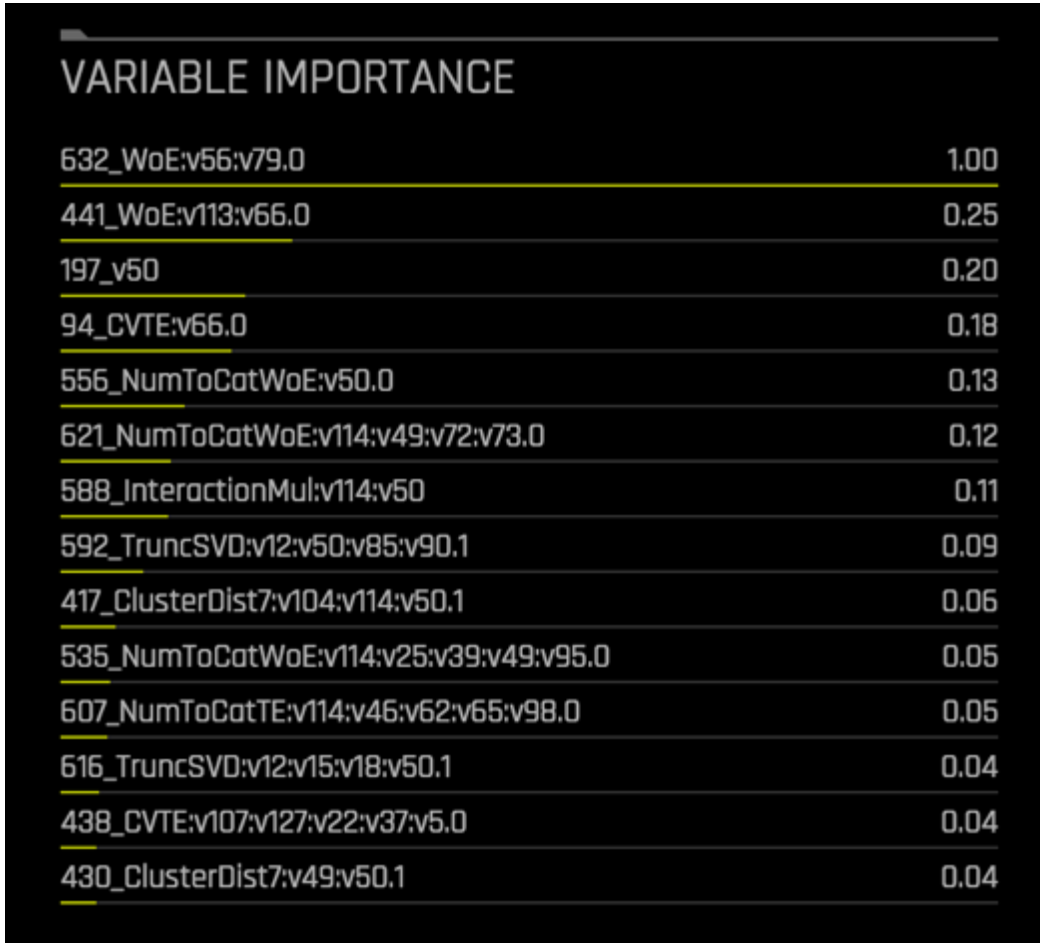
Automatic Visualization

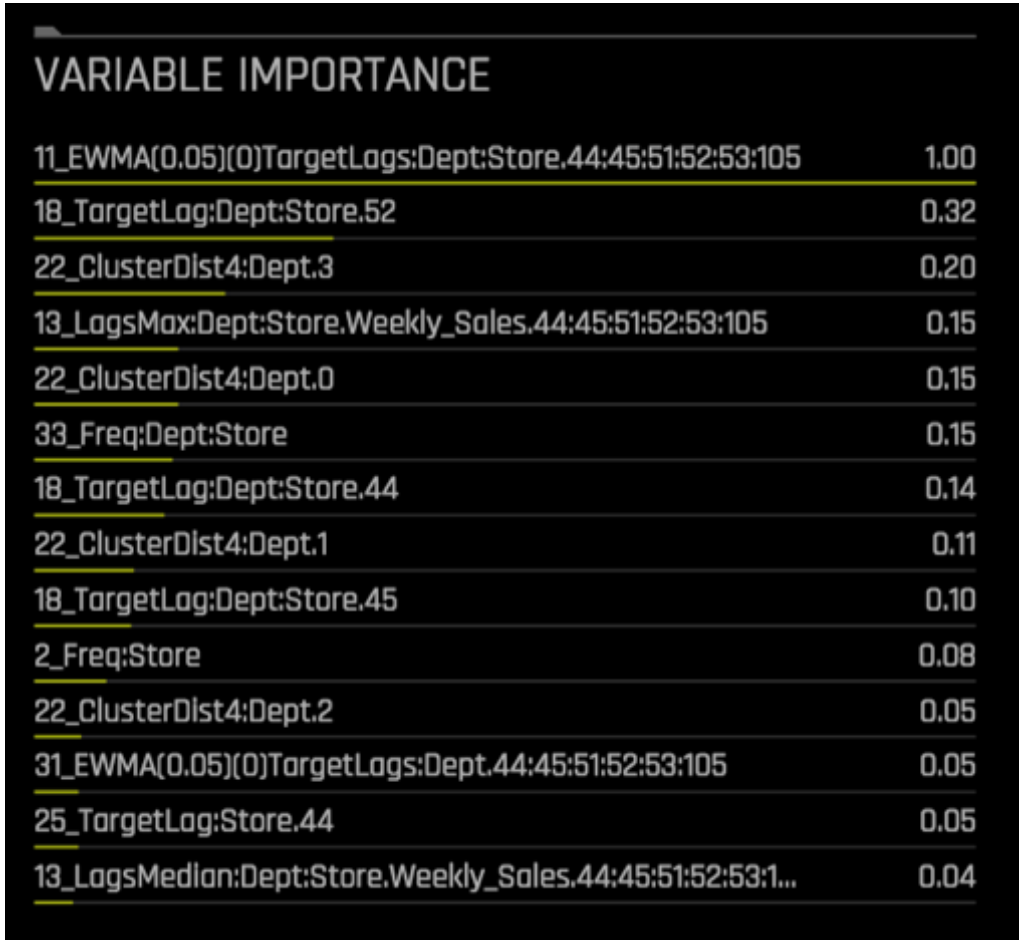
Automatic Scagnostics and other visualizations to generate the most relevant visualizations for each dataset





Secret Sauce: 1) Grandmaster Feature Engineering





Numerical/Categorical Interactions, Target Encoding, Clustering, Dimensionality Reduction, Weight of Evidence, etc.

Time-Series: Lags and historical aggregates with causality constraints

Secret Sauce: 2) Grandmaster Pipeline Tuning + Validation

Example: Driverless AI BNP Paribas on 3-GPU workstation

```
Recipe: AutoDL (171 iterations, 12 individuals)
    Validation scheme: stratified, 1 internal holdout
    Feature engineering: 18923 features tested (344 selected)
Timing:
    Data preparation: 8.44 secs
    Model parameter tuning: 403.98 secs (19 models trained)
    Feature engineering: 15424.53 secs (1008 models trained)
    Final model training: 1935.21 secs (26 models trained)
Validation score: LOGLOSS = 0.47811 +/- 0.0023019 (baseline)
Validation score: LOGLOSS = 0.43681 +/- 0.0037107 (final model)
Test score:       LOGLOSS = N/A (no target)
```

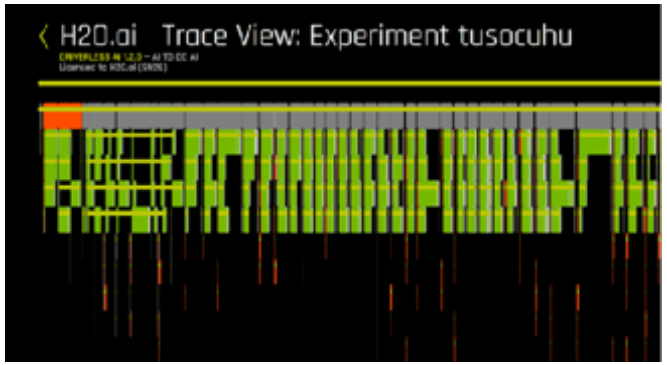
19,000 features tested

1,000 models trained

1 final optimal scoring pipeline

reliable generalization estimates (overfitting avoidance)

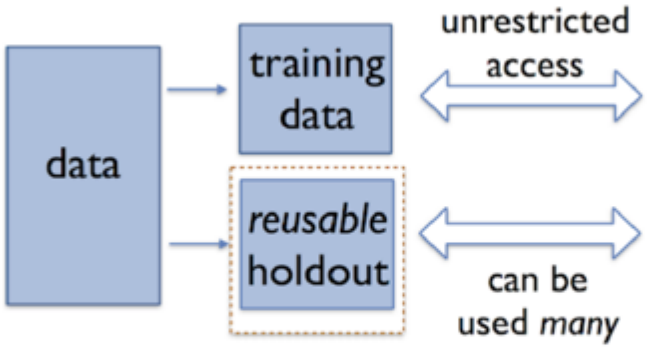
MTV



evolutionary strategies

massively parallel processing (multi-CPU, multi-GPU)

Reusable holdout method

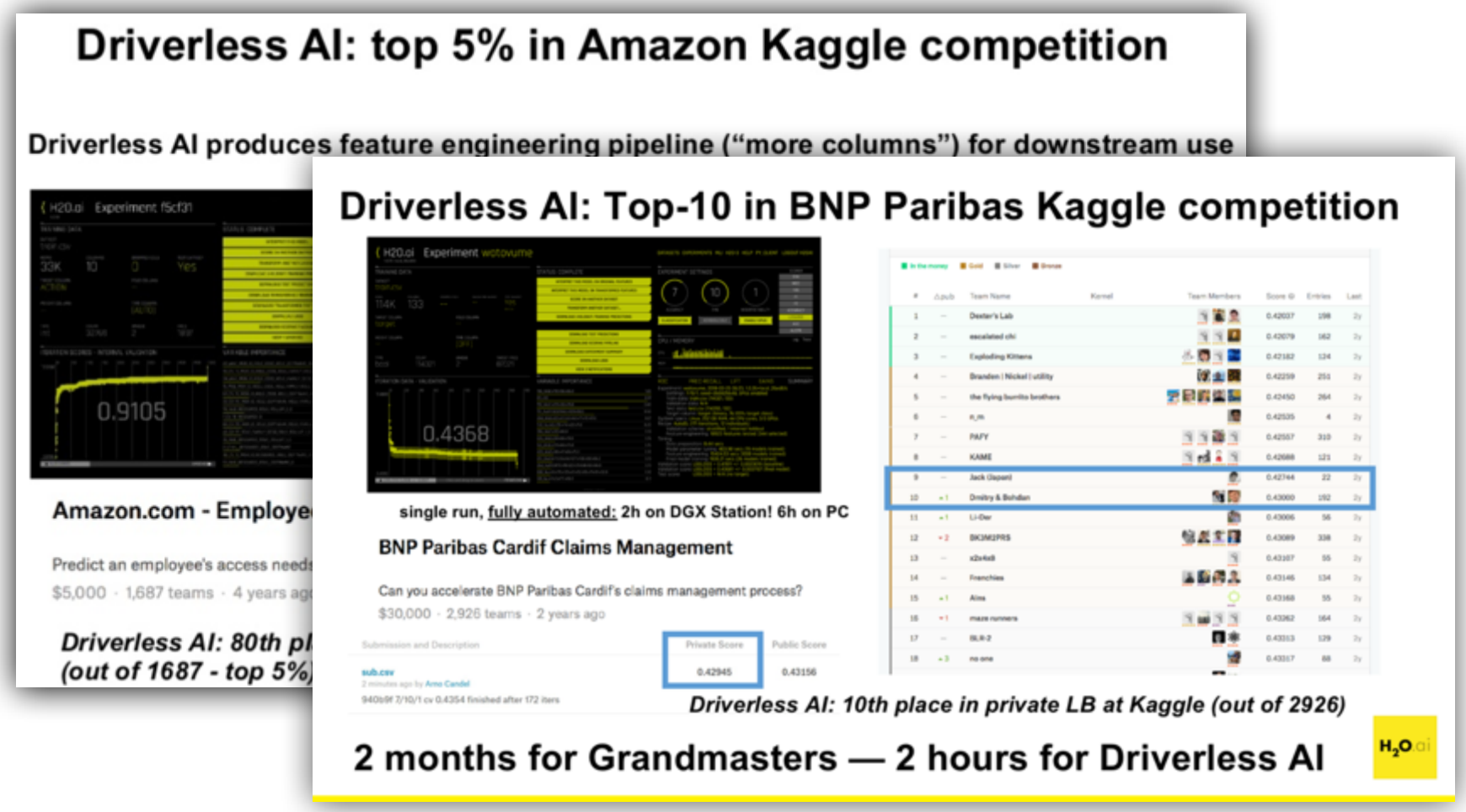


essentially as good as using *fresh* data each time!

DOI: 10.1126/science.aaa9375

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



Interpretability

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

Statistical Learning vs Deep Learning - We Do Both!

What do these settings mean?

ACCURACY
- Training data size: **1,000,000 rows, 8 cols**
- Feature evolution: **GBM, 5-fold CV**
- Final pipeline: **Ensemble (5xGBM, 2xTensorFlow), 5-fold CV**

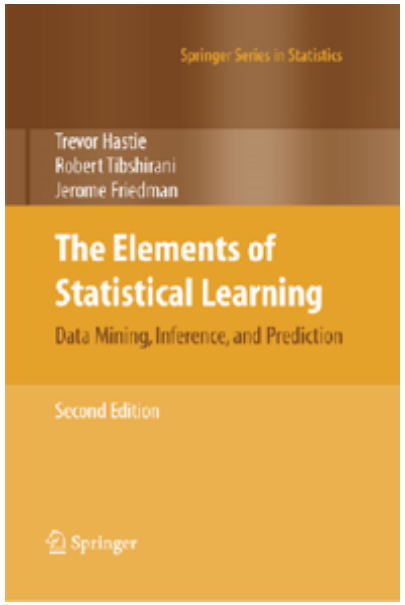
TIME
- Feature evolution: **8 individuals**, up to **47 iterations**
- Early stopping: After **5** iterations of no improvement

INTERPRETABILITY
- Feature pre-pruning strategy: None
- Monotonicity constraints: disabled
- Feature engineering search space (where applicable): ['Clustering', 'Date', 'FrequencyEncoding', 'Identity', 'Interactions', 'NumEncoding', 'TargetEncoding', 'Text', 'TruncatedSVD', 'WeightOfEvidence']

GBM models to train:
- Model and feature tuning: **165**
- Feature evolution: **620**
- Final pipeline: **35**

Estimated max. total memory usage:
- Feature engineering: **2.6GB**
- GPU XGBoost: **656.0MB**

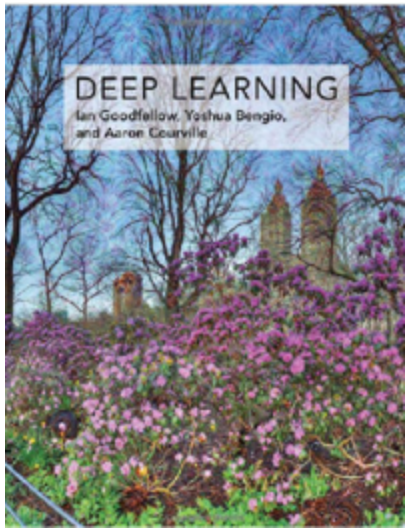
Estimated runtime: **10 minutes**



GLM/CART/RF/GBM/XGBoost
K-Means/PCA/SVD

Typically better for structured data (CSV, SQL, Transactional)

https://web.stanford.edu/~hastie/Papers/ESLII.pdf



TensorFlow Deep Learning

Typically better for unstructured data (Images, Video, Audio, Text)

http://www.deeplearningbook.org

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

Evaluation

Prizes

Timeline

About Bnp Paribas Cardif

As a global specialist in personal insurance, BNP Paribas Cardif serves 90 million clients in 36 countries across Europe, Asia and Latin America.

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.

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.

H2O.ai

Driverless AI Delivers "Expert Data Scientist in a Box"

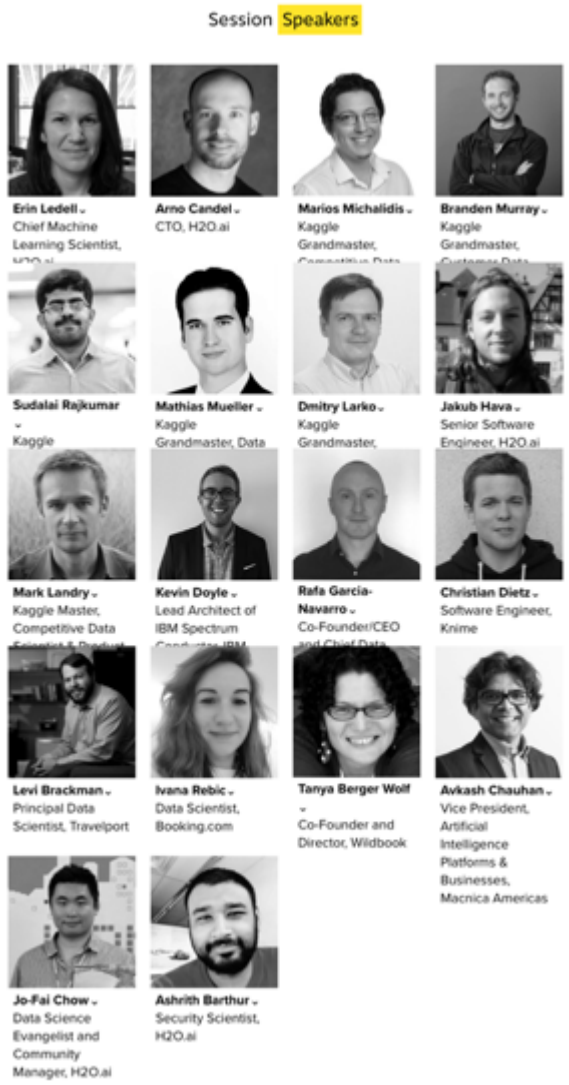
- Created and supported by world renowned AI experts
- Empowers companies to accomplish AI and ML with a single platform
- Performs the function of an expert data scientist and adds more power to both novice and expert teams
- Details and highlights insights and interpretability with easy to understand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

21 day free trial for Driverless AI

Our Flagship Community Event – **H2O AI World**

is finally coming to London!

29th & 30th Oct, London



More real-world use cases

+

All H_2O Kaggle Grandmasters

+

Hands-on Training

H_2O.ai

Thanks!

- More Info, Code, and Slides
  - bit.ly/ h2o_meetups
- Contact
  - joe@h2o.ai
  - @matlabulous
  - github.com/woobe