Speedup Learning

Faculty

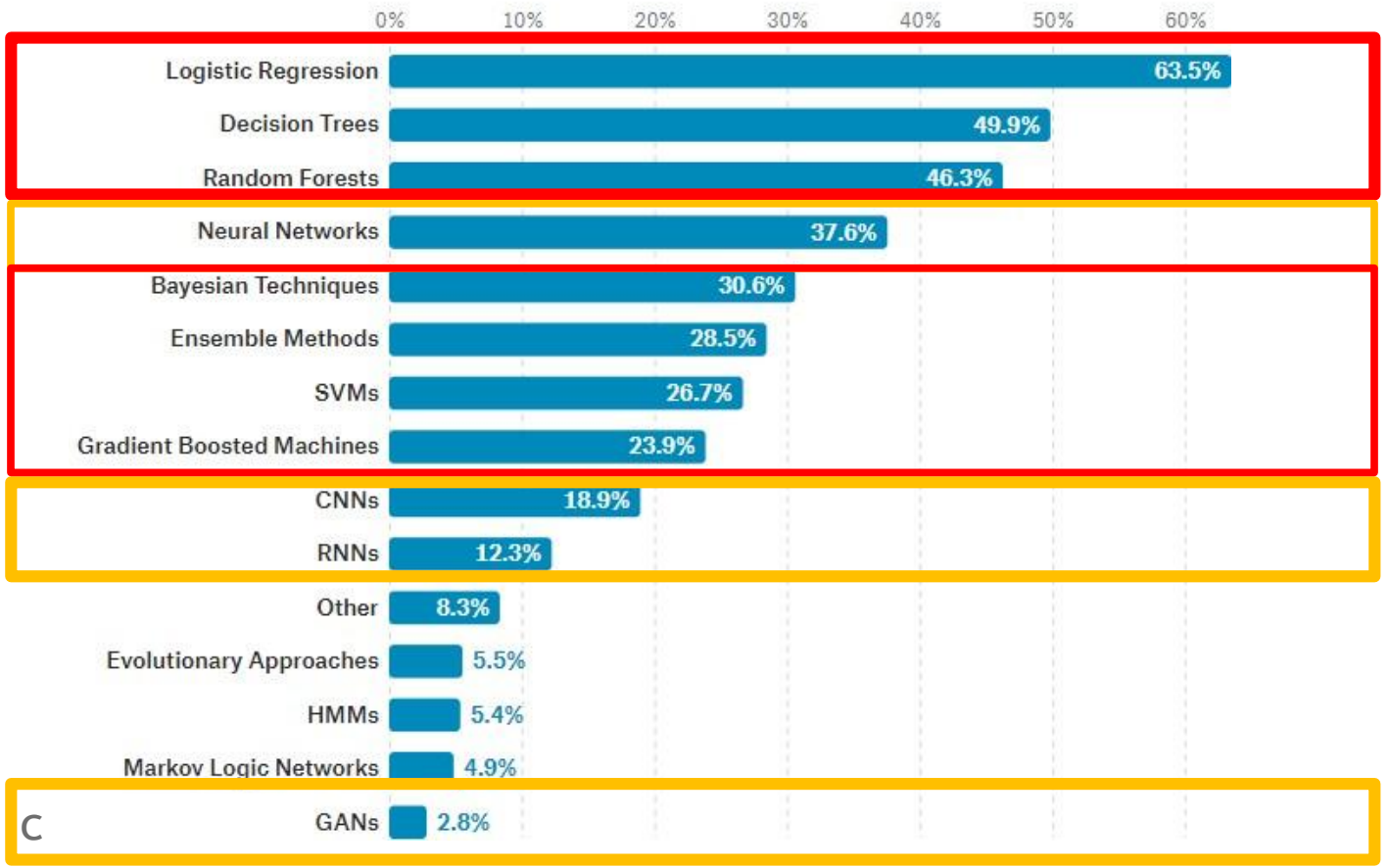
R. Rajkumar

School of Computing | SRMIST

What data science methods are used at work?

Logistic regression is the most commonly reported data science method used at work for all industries *except* Military and Security where Neural Networks are used slightly more frequently.

Company Size ▾ | Industry ▾ | Job Title ▾



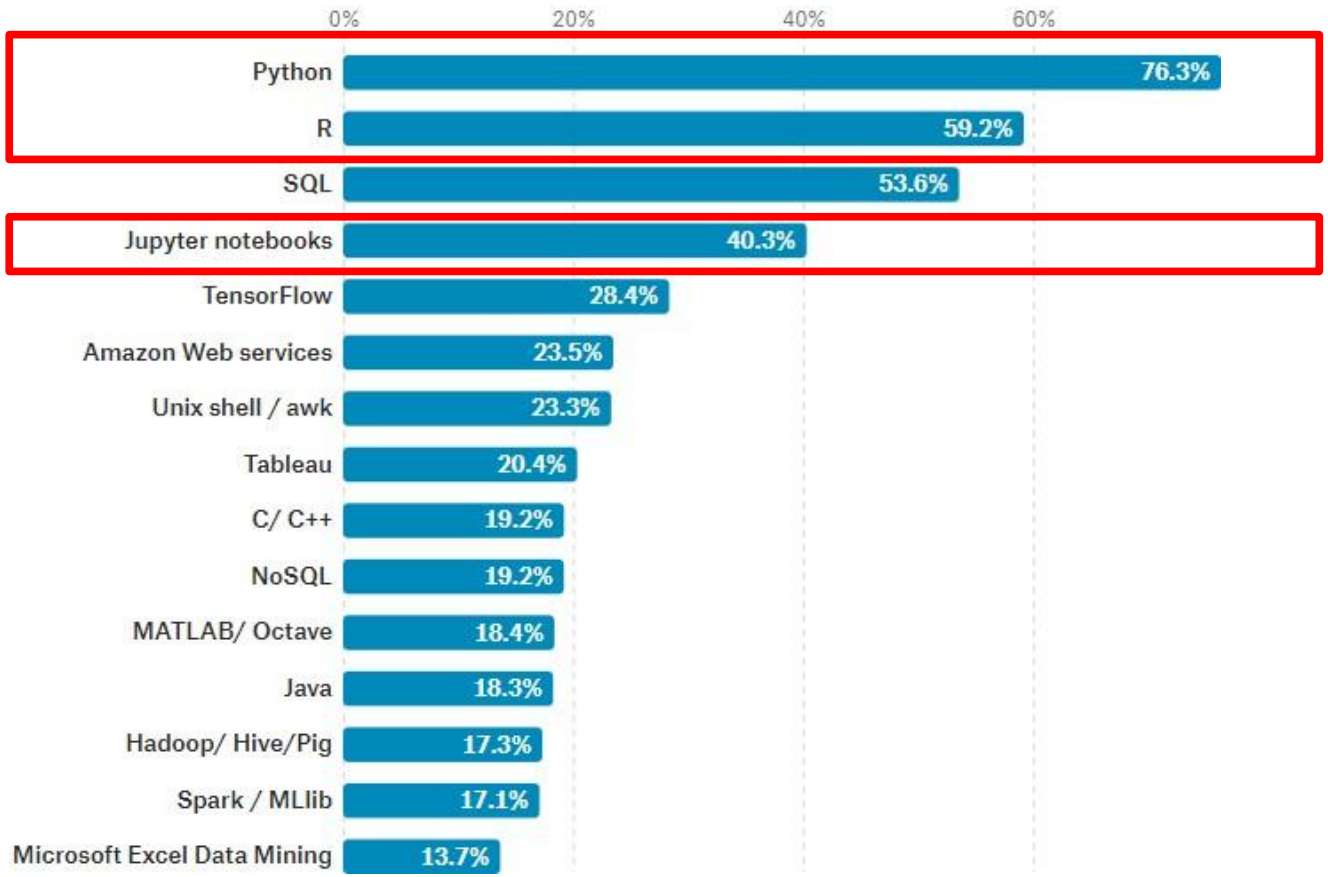
Deep Learning

Machine Learning

What tools are used at work?

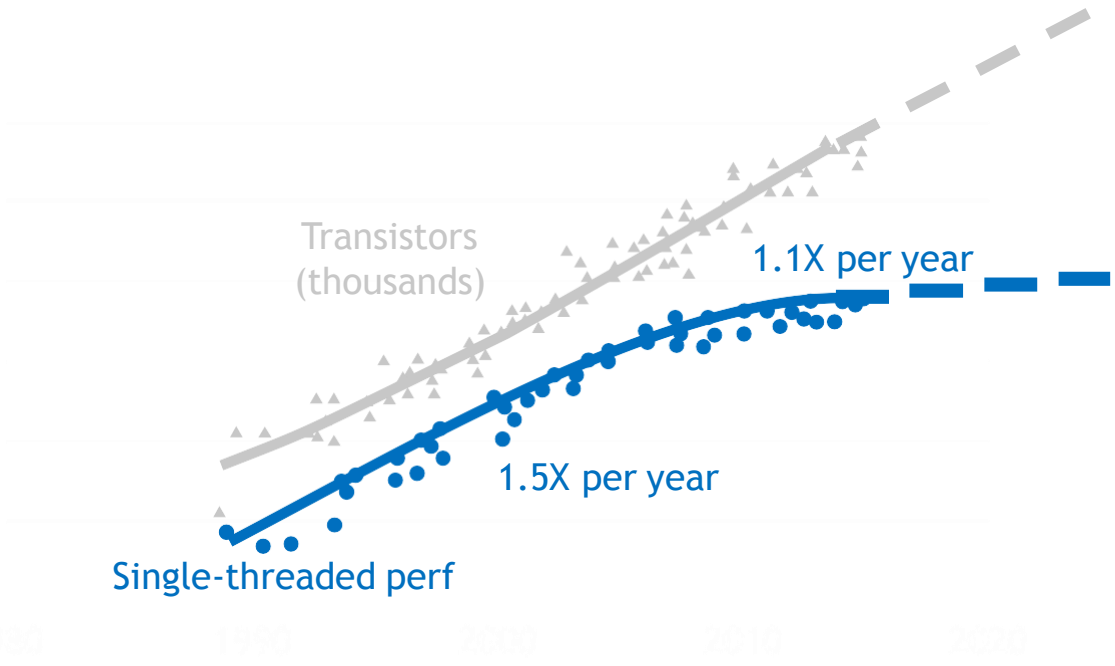
Python was the most commonly used data analysis tool across employed data scientists overall, but more Statisticians are still loyal to R.

Company Size ▼ Industry ▼ Job Title ▼



The End of Road for General Purpose Processors and the Future of Computing

John Hennessy
Stanford University
March 2017



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

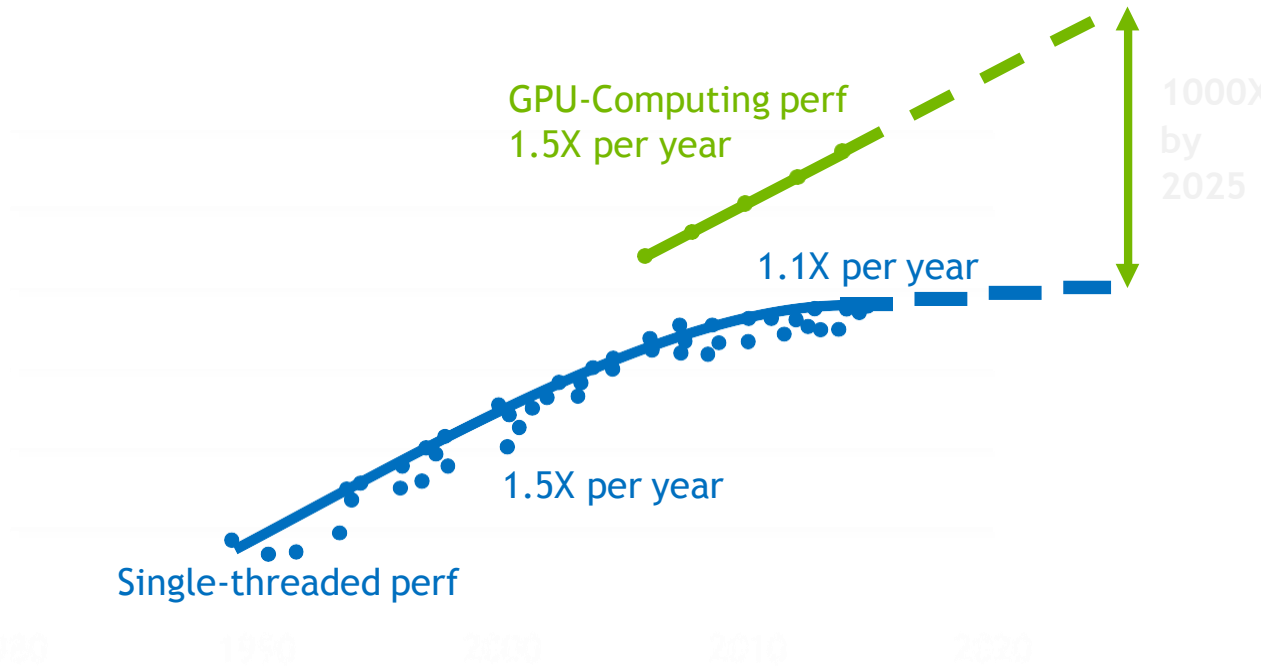
APPLICATIONS

ALGORITHMS

SYSTEMS

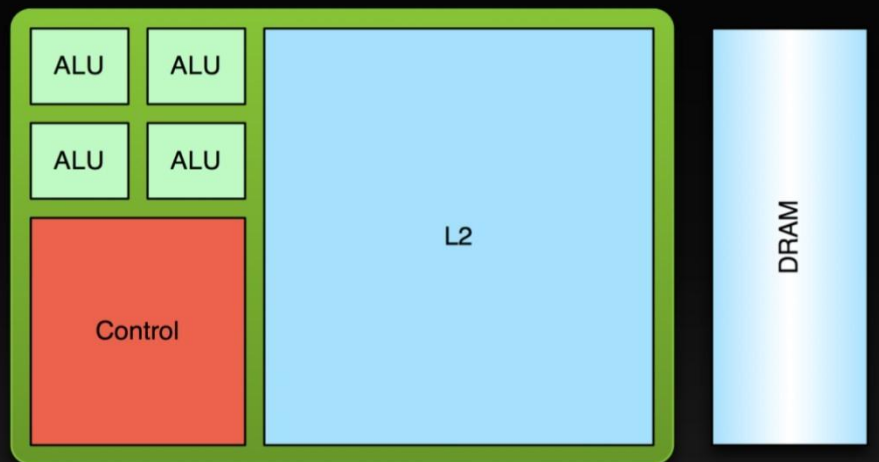
CUDA

ARCHITECTURE



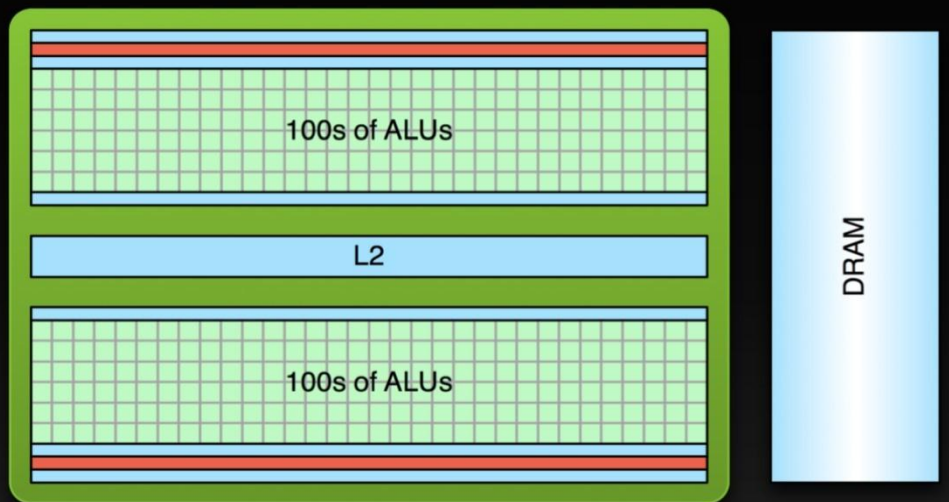
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

Low Latency or High Throughput?



CPU

- Optimized for low-latency access to cached data sets
- Control logic for out-of-order and speculative execution

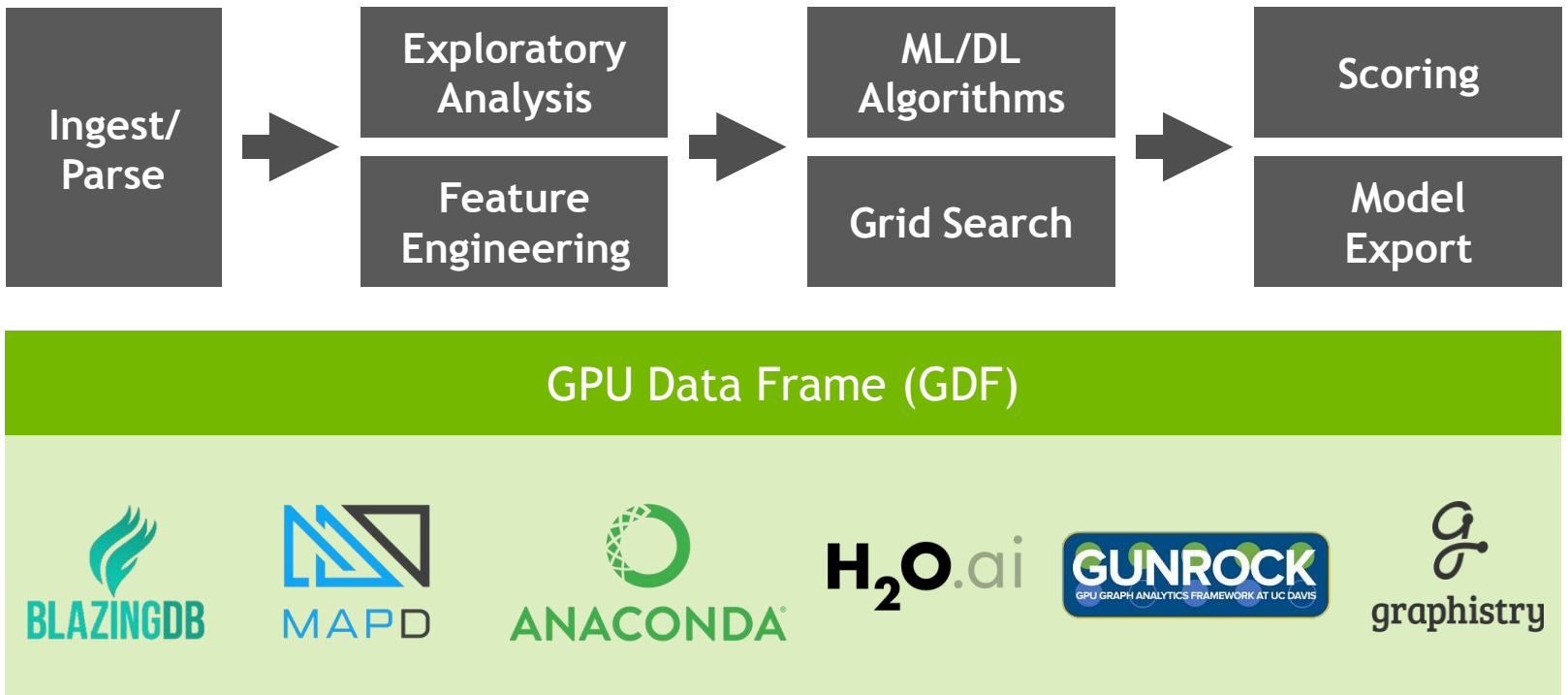


GPU

- Optimized for data-parallel, throughput computation
- Architecture tolerant of memory latency
- More transistors dedicated to computation

GPU OPEN ANALYTICS INITIATIVE

github.com/gpuopenanalytics



H2O4GPU

- / Used within our own **Driverless AI Product** to boost performance **30X**
  / Open-Source http://github.com/h2oai/h2o4gpu
- / Scikit-Learn Python API (and soon R API)
- / All Scikit-Learn algorithms included, important ones ported to GPU

H2O4GPU Roadmap

Currently Available - Q3 (09-30-2017)
GLM (POGS)
Python API for training & scoring
GBM
Inference on GPU (GLM)
Random Forest
Inference on GPU (GBM)
k-Means Clustering

API Support
Python API for training & scoring
Sckit learn API compatibility

Q4 2017 - (12-31-2017)
k-Nearest Neighbors
PCA
SVD
Quantiles
Kalman Filters
Sort
Aggregator

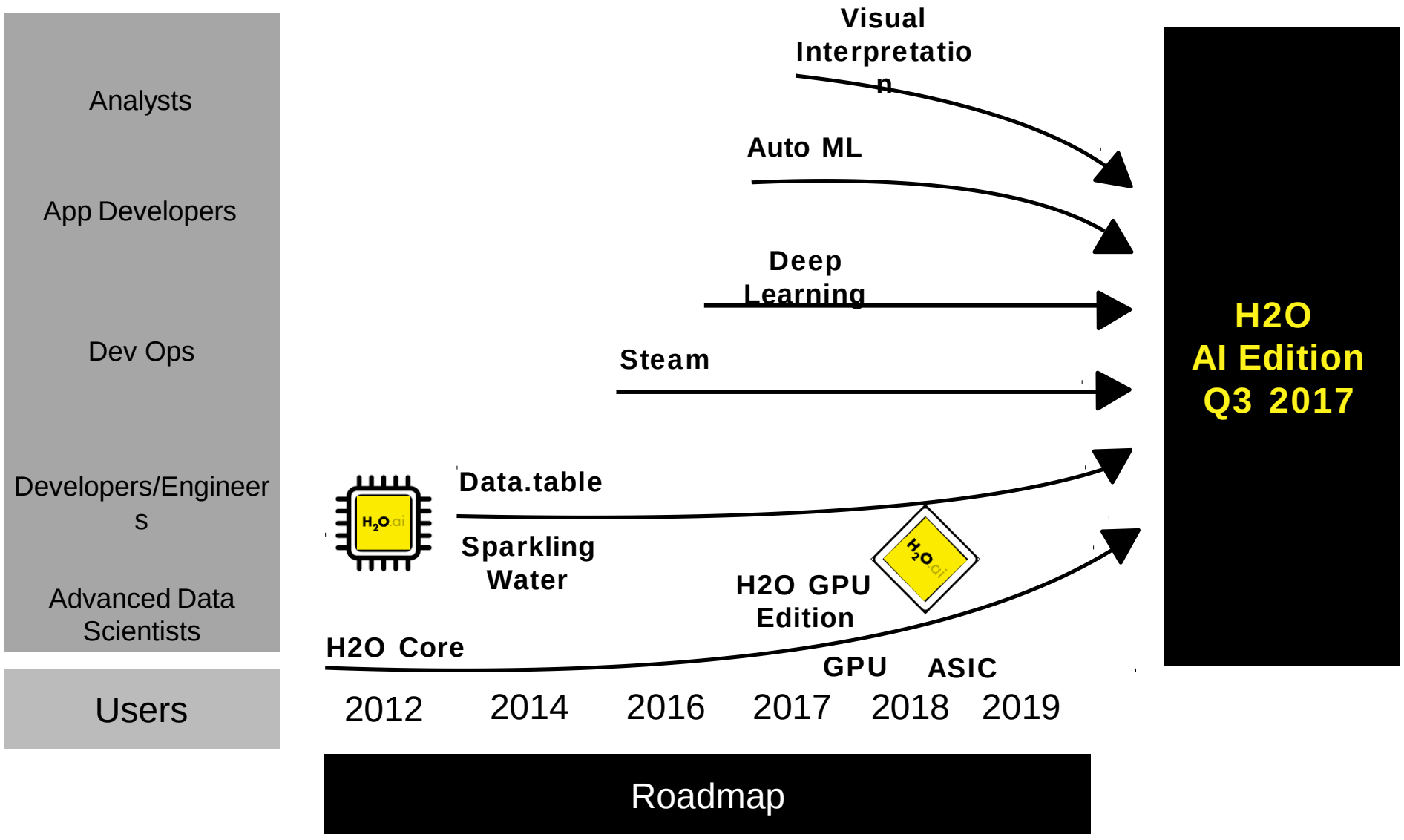
API Support
R API for training & scoring
GOAI API support
Data.table

Performance & Scalability
Fastest single GPU performance
Multi GPU
Multi machine

2018-19
Kernel Methods
Recommendation Engines - Non-Negative Matrix Factorization
Recommendation Engines - Bayesian Neural Nets
MCMC Solver
Time Series
SVM
Text Analysis - TF-IDF
Text Analysis - Word2Vec
Text Analysis - Doc2Vec
Automatic K for K-means
H2O GLM - Lasso
Simulation Techniques
Sampling Techniques

Domain Specific Algorithms
Life Sciences
Financial Services
Underwriting
Sampling Techniques

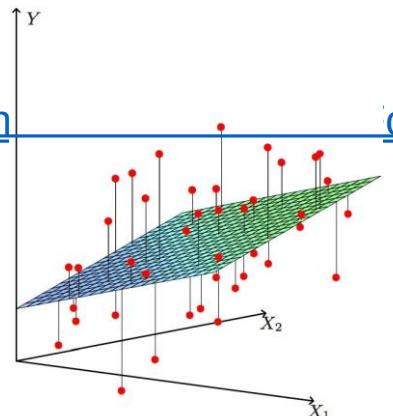
H2O AI Platform Timeline

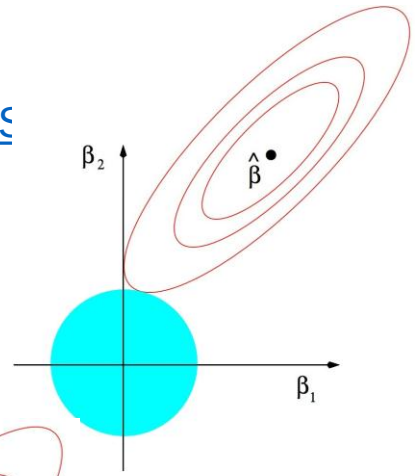


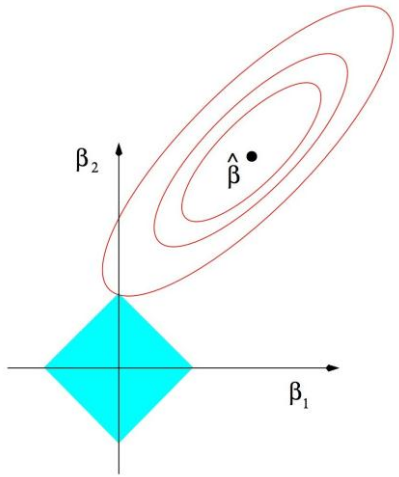
Model
Accuracy & Speed

Generalized Linear Modelling in H2O4GPU

- Framework utilizes Proximal Graph Solver (POGS) from Stephen Boyd & Chris Fougner (Parameter Selection and Pre-Conditioning for a Gra ph ougner and S)
  - A solver for convex optimization problems in *graph form* using Alternating Direction Method of Multipliers (ADMM)
- Solvers include Lasso, Ridge Regression, Logistic Regression, and Elastic Net Regularization
- Improvements to original implementation of POGS:
  - Full alpha search
  - Cross Validation
  - Early Stopping (RMSE for regression problems and Logloss for classification)
  - Various bug fixes from original implementation
  - Added Scikit learn “like” API
  - Supports multiple GPUS





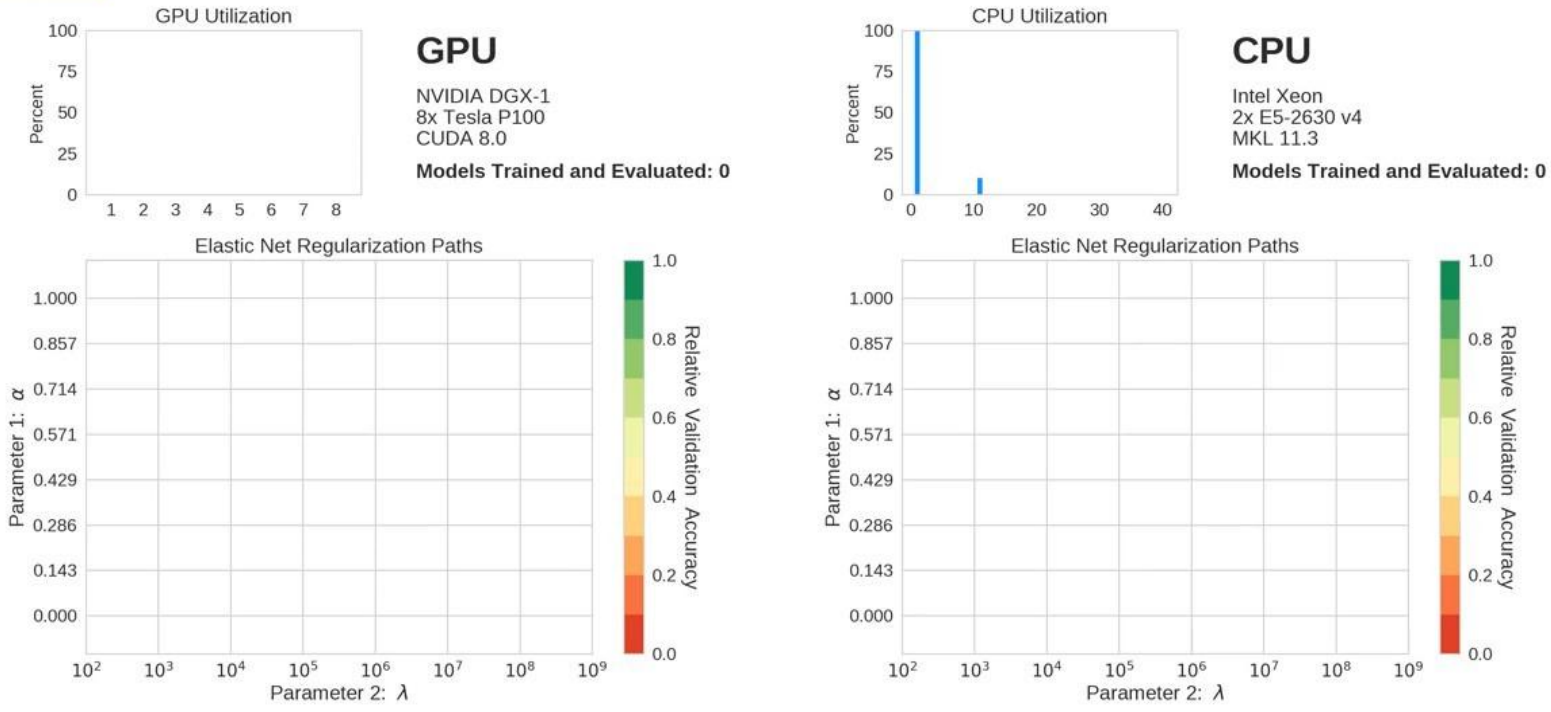


Hastie, Tibshirani, Friedman. Elements of Statistical Learning

Hastie, Tibshirani, Wainwright. Statistical Learning with Sparsity

YouTube

H2O.ai Machine Learning – Generalized Linear Modeling



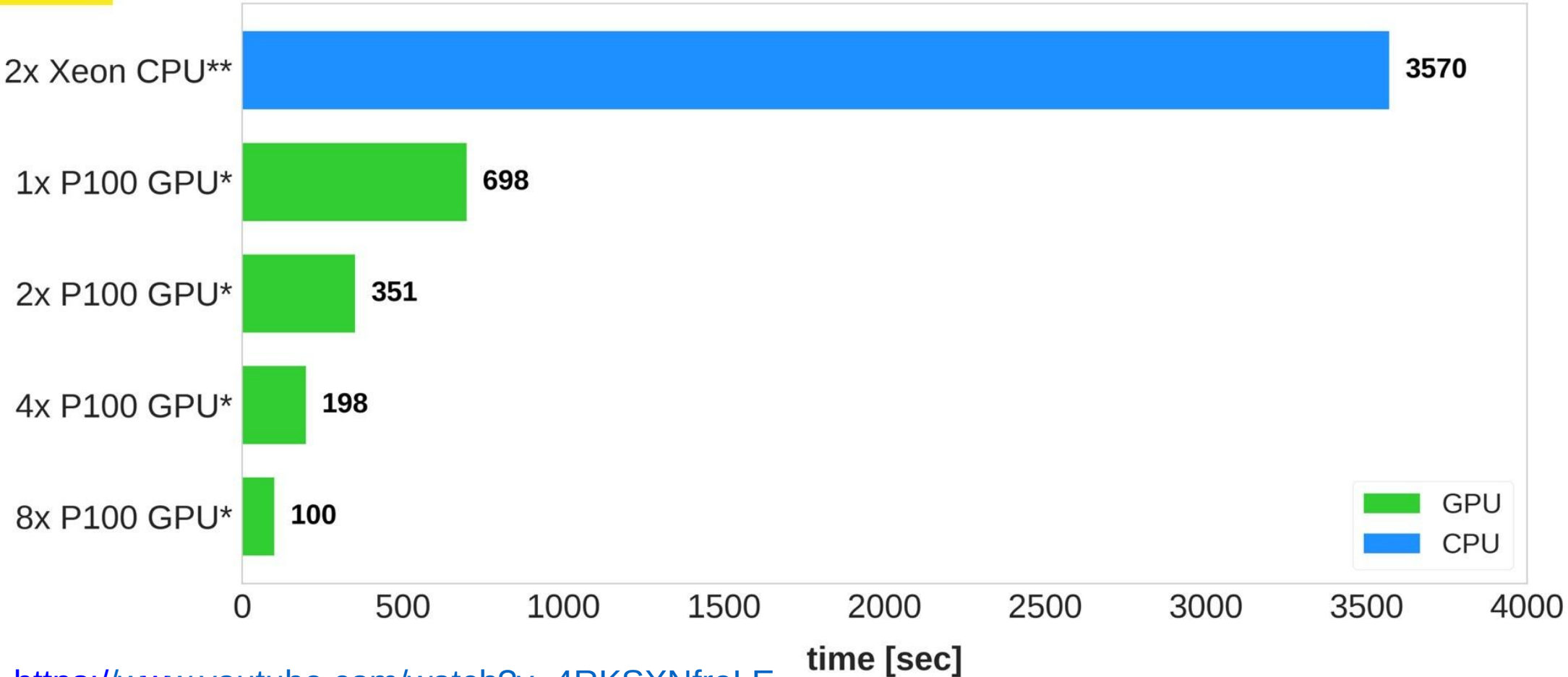
U.S. Census dataset (predict Income): 45k rows, 10k cols
Parameters: 5-fold cross-validation, $\alpha = \{\frac{i}{7}, i = 0 \ldots 7\}$, full λ-search

https://www.youtube.com/watch?v=LrC3mBNG7WU

https://github.com/h2oai/h2o4gpu/blob/master/exam

H2O.ai Machine Learning – Generalized Linear Modeling

Time to Train and Evaluate 4000 Models



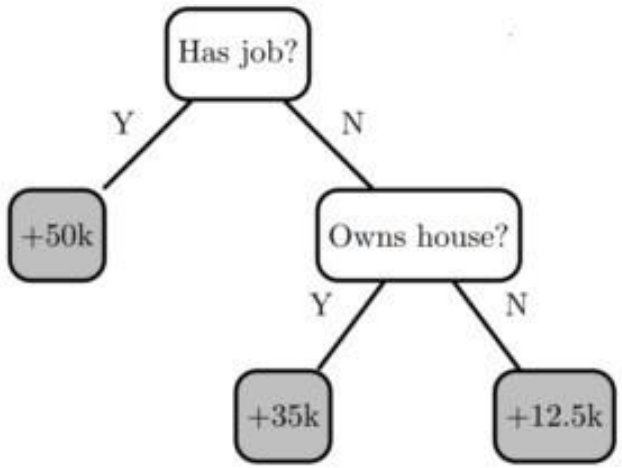
https://www.youtube.com/watch?v=4RKSXNfreLE

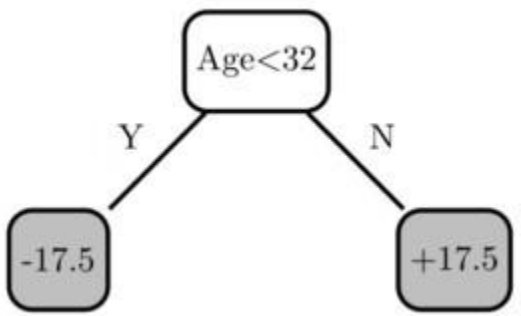
http://github.com/h2oai/perf/

*NVIDIA DGX-1, **Dual Intel Xeon E5-2630 v4
U.S. Census dataset (predict Income): 45k rows, 10k cols
Elastic Net Model Parameters: 5-fold cross-validation, $\alpha = \{\frac{i}{7}, i = 0 \ldots 7\}$, full λ-search

Gradient Boosting Machines in H2O4gpu

- Based upon XGBoost
- Raw floating point data -> Binned into Quantiles
- Quantiles are stored as compressed instead of floats
- Compressed Quantiles are efficiently transferred to GPU
- Sparsity is handled directly with highly GPU efficiency
- Multi-GPU by sharding rows using NVIDIA NCCLAllReduce

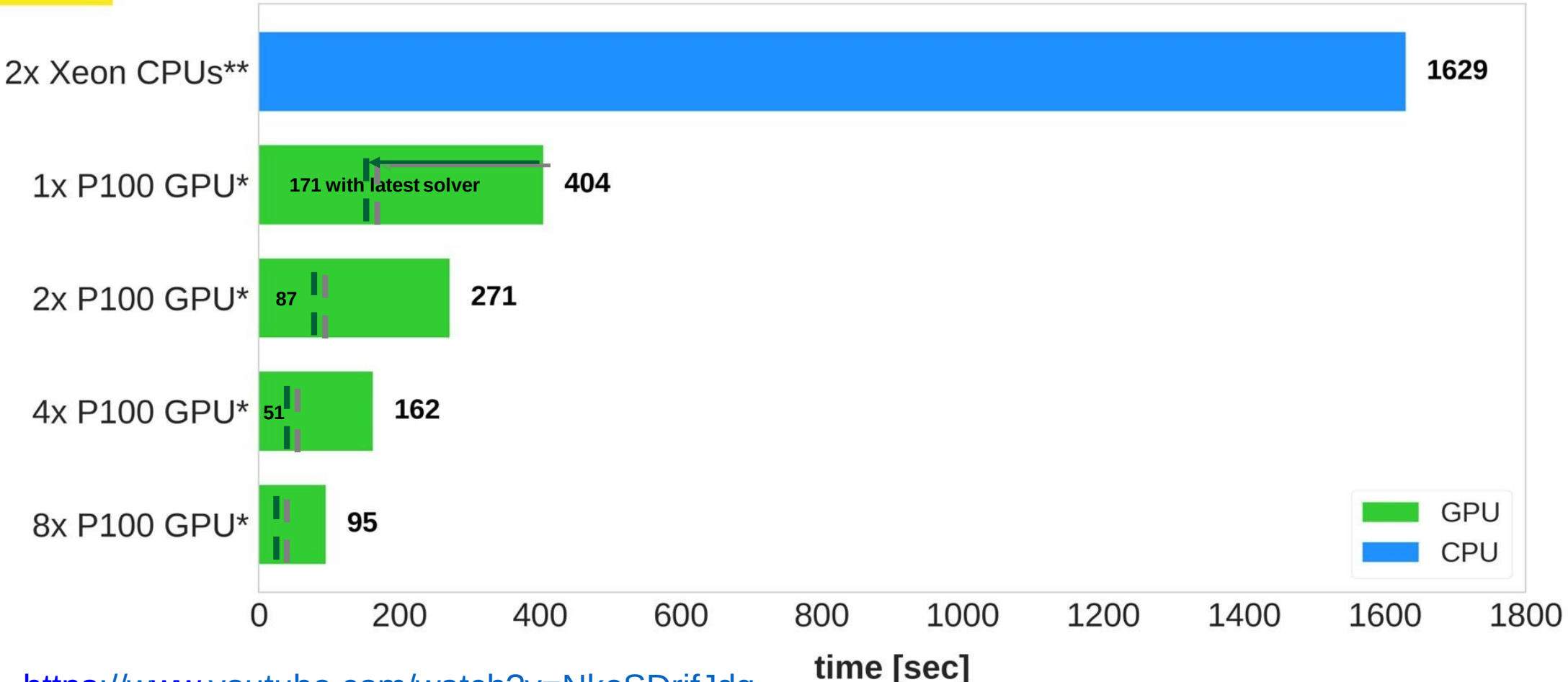




https://github.com/h2oai/h2o4gpu/blob/master/examples/py/xgboost_simple_demo.ipynb

H2O.ai Machine Learning – Gradient Boosting Machine

Time to Train 16 H2O XGBoost Models

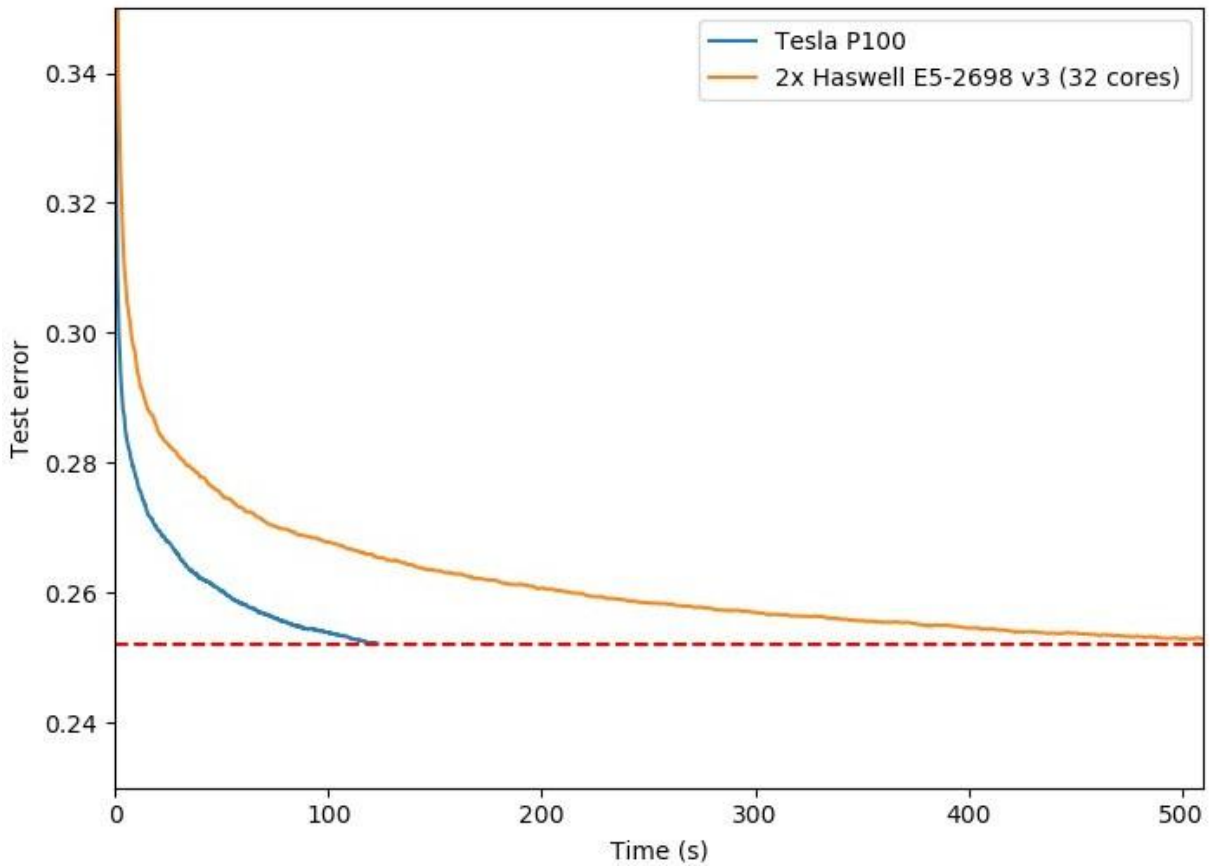


https://www.youtube.com/watch?v=NkeSDrifJdg

*NVIDIA DGX-1, **Dual Intel Xeon E5-2630 v4

http://github.com/h2oai/perf/ Higgs dataset (binary classification): 1M rows, 29 cols; max_depth: {6,8,10,12}, sample_rate: {0.7,0.8,0.9,1.0}

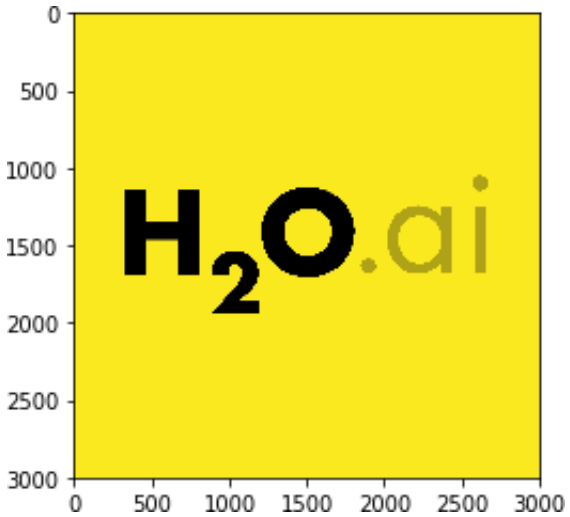
CPU vs. GPU on Higgs (Classification)

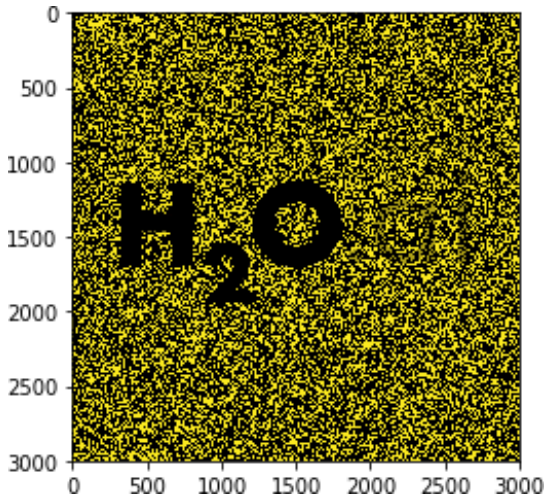


K-Means on H2O4gpu

- Based upon NVIDIA prototype of K-Means algorithm in CUDA
- Improvements to original implementation:
  - Significantly faster than scikit-learn implementation (50x)
  - Significantly faster than other GPU implementations (https://github.com/src-d/kmcuda) (5x-10x)
  - Various bug fixes
  - Supports multiple GPUs

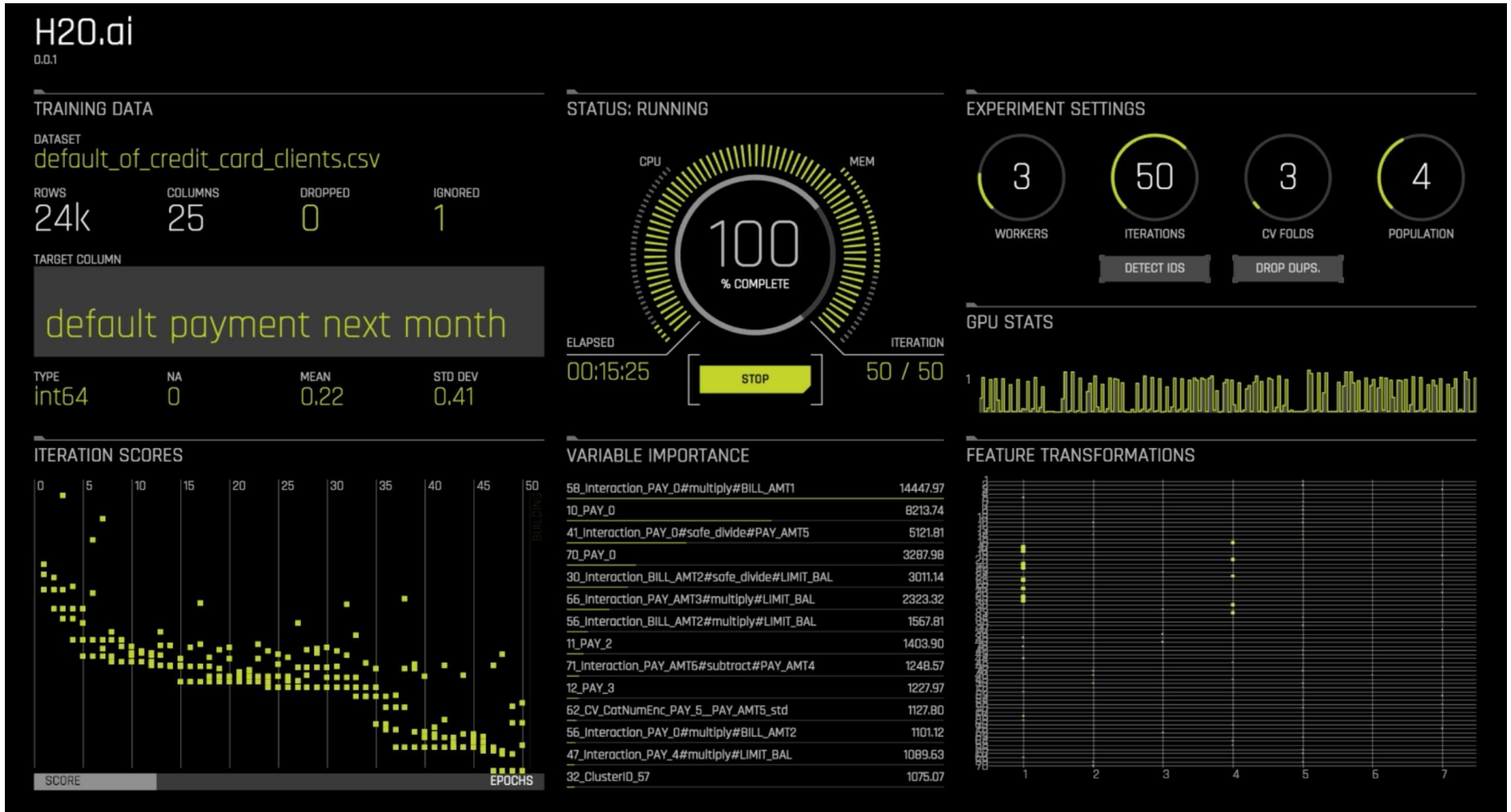
K-Means



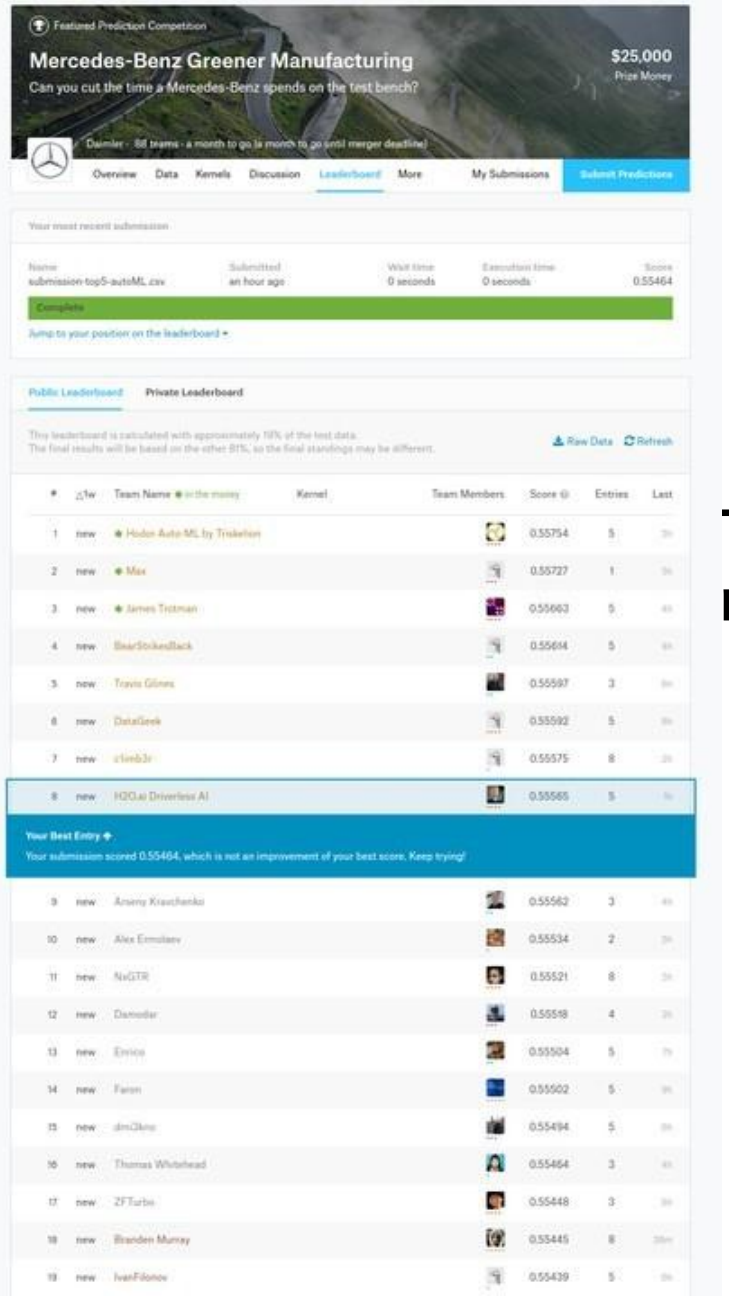


https://github.com/h2oai/h2o4gpu/blob/master/examples/py/demos/H2O4GPU_KMeans_Images.ipynb

Driverless AI on GPUs



https://www.youtube.com/watch?v=KkvWX3FD7yI



Driverless AI — Competitive with Kagglers!

'op 8 position in Kaggle with zero manual labor!
ranked above multiple Kaggle Grandmasters)

https://www.kaggle.com/c/mercedes-benz-greener-manufacturing/leaderboard